AD_____

Award Number: W81XWH-05-1-0031

TITLE: High-Resolution Mapping of Structural Mutations in Prostate Cancer with Single Nucleotide Polymorphism Arrays

PRINCIPAL INVESTIGATOR: Rameen Beroukhim, Ph.D.

CONTRACTING ORGANIZATION: Dana Farber Cancer Institute
Boston, MA 02115

REPORT DATE: November 2005

TYPE OF REPORT: Annual Summary

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-11-2005	Annual Summary	1 Nov 2004 – 31 Oct 2005

4. TITLE AND SUBTITLE
High-Resolution Mapping of Structural Mutations in Prostate Cancer with Single Nucleotide Polymorphism Arrays

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0031

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Rameen Beroukhim, Ph.D.

E-mail: rberoukhim@partners.org

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Dana Farber Cancer Institute
Boston, MA 02115

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The proposal focused on the systematic mapping of large-scale genetic alterations in prostate cancer, and relating these mutations to prostate cancer progression. To that end, the proposal suggested the application of single nucleotide polymorphism (SNP) array technology to characterize large-scale genetic alterations in the prostate cancer genome. In the 11 months since the beginning of the award, significant progress has been made in all 3 specific aims. Multiple primary and metastatic prostate tumors have been collected and genome-wide maps loss of heterozygosity and copy number changes have been generated from 100K SNP array data. Regions of significant chromosomal aberrations have been identified, and in this preliminary analysis several of these aberrations have been found to correlate with prostate cancer progression. Unanticipated difficulties have arisen with laser capture microdissection of primary prostate cancers, however, and efforts are ongoing to surmount these difficulties. During the conduct of this research, a method for determination of loss of heterozygosity without the use of paired normals, was developed to account for the haplotype structure of the genome, and a manuscript describing this method is in review.

15. SUBJECT TERMS
Prostate Cancer, Genomics, Chromosome Structure, Cancer Progression and Metastasis, Single Nucleotide Polymorphisms, Genotyping, Digital Karyotyping, Cytogenetics, Loss of Heterozygosity, Oligonucleotide Array

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					USAMRMC
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	50	

Table of Contents

A. Introduction

The proposal for DOD Award #W81XWH-05-1-0031 focused on the systematic mapping of large-scale genetic alterations in prostate cancer, and relating these mutations to prostate cancer progression. To that end, the proposal suggested the application of single nucleotide polymorphism (SNP) array technology to characterize large-scale genetic alterations in the prostate cancer genome.

1. High-resolution single nucleotide polymorphism arrays

SNPs are the most common genetic variation in the human genome; more than 6,000,000 have been identified (Sachidanandam et al., 2001). The use of single-nucleotide polymorphisms to study the germline genetic susceptibility to disease is well appreciated and an evolving technology designed to conduct such studies is the use of oligonucleotide arrays to interrogate these SNP markers in a high-throughput, highly parallel fashion (Cutler et al., 2001; Matsuzaki et al., 2004; Matsuzaki et al., 2004). These oligonucleotide arrays specifically detect the two different alleles of each SNP (Figure 1). The most advanced commercially available 100K arrays detect 116,204 SNPs. With a median intermarker distance of 8.9 kb, this represents greater than 5 SNPs per gene, affording state-of-the art resolution for large-scale genotyping purposes (Craig and Stephan, 2005).

Figure 1: View of a probe set for a single SNP showing a homozygous "A" call. For each SNP, 40 oligonucleotide probes are tiled onto SNP arrays interrogating 116,204 SNPs across the genome. These include perfect match (pm) and mismatch (mm) probes directed against each allele. The SNP position slides 5' to 3' among the pm or mm probes to each allele. The fluorescence pattern indicates which alleles are present; the fluorescence intensity indicates the quantity of bound DNA.

To prepare target for the 100K arrays, genomic DNA is digested with XbaI or HindIII (in separate reactions). HindIII and XbaI linkers are ligated and single-primer PCR amplification is carried out to amplify fragments ranging from 200-2000 bp, resulting in a partial genome representation. The fragments are labeled with streptavidin, fragmented and hybridized to arrays that contain the 58,000 probe sets for either the XbaI or HindIII digest. The probe set for each SNP consists of 10 perfect match (pm) probes to each allele, along with 10 mismatch (mm) probes, for a total of 40 probes. A detailed description of the protocols and technology for these 100K SNP arrays is available at www.affymetrix.com/support/technical/datasheets/100k_datasheet.pdf.

The scale and precision with which high-density SNP arrays interrogate independent alleles prompted our group (led by William Sellers and Matthew Meyerson) to spearhead efforts applying this technology to the analysis of somatic genetic alterations present in human malignancies. Several features of SNP arrays suggested that they might constitute an ideal platform for cancer genomic analyses: 1, determination of allele status across cancer genomes provided a basis for large-scale, high-precision loss of heterozygosity (LOH) analysis; 2, probe set hybridization yielded a signal whose intensity also reflected the copy number at that locus; and 3, the resolution afforded by SNP array marker densities exceeded that of most CGH options.

a. High-resolution loss of heterozygosity analysis

The somatic conversion of heterozygous germline alleles to a homozygous state (LOH) may occur through hemizygous deletion alone (resulting in concomitant copy loss) or followed by gene duplication (copy-neutral LOH). Interestingly, copy-neutral LOH, which is undetectable by conventional CGH methods, represents up to 80% of LOH events in some tumor sets (Huang et al., 2004), and the primary mechanism of LOH in particular genomic regions of individual cancer types (Irving et al., 2005; Jacqueline A. Langdon, 2005). Considerable experimental evidence supports the notion that LOH represents a key mechanism for tumor suppressor inactivation. Indeed, nearly all common tumor suppressor genes occur in regions that frequently undergo LOH (prominent examples include p16, PTEN, pRB, and p53).

Published data by the Sellers and Meyerson groups and by others demonstrate that SNP arrays provide high-resolution maps of LOH when one compares the pattern of heterozygosity in the constitutional germline DNA to the pattern seen in the tumor (Allinen et al., 2004; Dumur et al., 2003; Hoque et al., 2003; Janne et al., 2004; Lieberfarb et al., 2003; Lindblad-Toh et al., 2000; Mei et al., 2000; Paez et al., 2004; Primdahl et al., 2002; Schubert et al., 2002; Wang et al., 2004). More recently, we have developed methods of analyzing homozygous allele frequencies and regions of linkage

disequilibrium to map regions of LOH without the use of paired normal germline DNA samples (Beroukhim et. al., in review and Body, below). This has allowed us to map LOH in cell lines and xenografts and to determine the similarity or differences in this data compared to authentic human tumors.

b. Genome-wide maps of copy number aberrations

Our group and others have found that comparison of signal intensities derived from each SNP probe (instead of allele call data) to corresponding signal data from normal genomes allows determination of copy number changes present within tumor samples (Bignell et al., 2004; Zhao et al., 2004). The concordance with quantitative PCR has generally been excellent, though high-level copy number gains are often underestimated on SNP arrays, presumably due to saturation effects. Various cancer genomes are now beginning to be mapped in this way (Rubin et al., 2004; Zhou et al., 2004), including analyses by our group, using 100K arrays, of the lung cancer genome (Zhao et al., 2005) and of the NCI60 cell line set (Garraway et al., 2005). The high resolution of the 100K arrays allowed the discovery, in this latter case, of the novel oncogene MITF in melanoma cell lines and metastatic samples.

To that end, the specific aims proposed were:

1. **To isolate DNA from 50 localized and 50 metastatic prostate cancers after laser-capture microdissection, along with DNA from corresponding germline tissue.**
2. **To generate genome-wide high-resolution maps of LOH and copy-number alterations using SNP arrays containing probes for 100,000 markers.**
3. **To identify and validate candidate somatic genetic alterations differing in prevalence between localized and metastatic cancers, and develop markers for clinical association studies.**

In the 11 months since the beginning of the award, significant progress has been made in all 3 specific aims. However, unanticipated difficulties have also arisen with respect to Specific Aim 1. The progress and difficulties will be outlined in the next section.

B. Body

1. Specific aim #1: To isolate DNA from 50 localized and 50 metastatic prostate cancers after laser-capture microdissection, along with DNA from corresponding germline tissue.

Reconstitution experiments have shown (Lindblad-Toh et al., 2000) that contamination of cancer cells with greater than 10% normal cell genomes results in a significant degradation in the ability to determine LOH. Prostate cancers tend to have large concentrations of intervening stroma. Thus, to apply SNP array technology to the study of prostate cancer, samples must be enriched for tumor. In this aim, we attempt to preserve the detection of both LOH events and copy number changes in prostate samples using laser capture microdissection (LCM)-based methods for tumor enrichment.

Our experience has shown that LCM of 2 mm^2 of prostate tissue takes between 2-4 hours and yields 50-100 ng of DNA. A 100K SNP array set requires 500 ng of DNA; to produce this amount of DNA on a large number of tumors quickly becomes prohibitively time-consuming. Fortunately, several methods of whole genome amplification (WGA) exist (Hughes et al., 2005). Among the most promising of these is multiple displacement amplification (MDA) (Dean et al., 2002), which makes use of a polymerase with exonuclease activity and random primers to perform isothermal amplification, with yields as high as 10,000-fold or greater. As opposed to PCR-based methods, the DNA produced has long fragment lengths and low error rates. We have shown (Paez et al., 2004) that, using 10 ng of high-quality template DNA, one can produce tens of micrograms of DNA with MDA methods using the Φ29 polymerase. The DNA product preserves genotyping information with 99.8% accuracy, and copy numbers determined from this DNA are 87% concordant with the unamplified template DNA. Much of the 13% discordance in copy number estimates was not functionally important, as it was due to lower saturation levels in WGA DNA—meaning very high amplifications (copy number 6 or greater) were not seen to be as high in WGA

as unamplified DNA—although they were noted to be high in both groups. Therefore, we have attempted to apply MDA to DNA obtained from laser-capture mcirodissected tissue

This section will describe progress in 3 sub-aims:

a. Characterization of LCM and WGA conditions for optimal reproducibility
b. Collection of primary and metastatic tumors for microdissection
c. Production of high-quality DNA from whole genome amplification of DNA obtained from laser-capture microdissected tumors and germline tissue

a. Characterization of LCM and WGA conditions for optimal reproducibility

Early on, we found that DNA from LCM can also serve as a template for WGA, providing product that gives similar results on SNP arrays to unamplified DNA. Among four highly enriched tumors, high-density SNP array data was obtained after either macrodissection (either a 2 mm cubic biopsy of tissue, or tissue needle-dissected from a glass slide) or laser capture microdissection. Overall call rates, reflecting the percentage of SNPs for which genotypes could be assigned, averaged 93.7% and 93.6% for macrodissected and microdissected tissue, respectively. Moreover, concordance rates between genotype calls from macrodissected and microdissected tissue averaged 98.2%. Although this concordance rate is slightly worse than that obtained with the highest quality DNA (99.85% in our hands (Paez et al., 2004)), it is high enough to accurately assign regions of LOH (Figure 2).

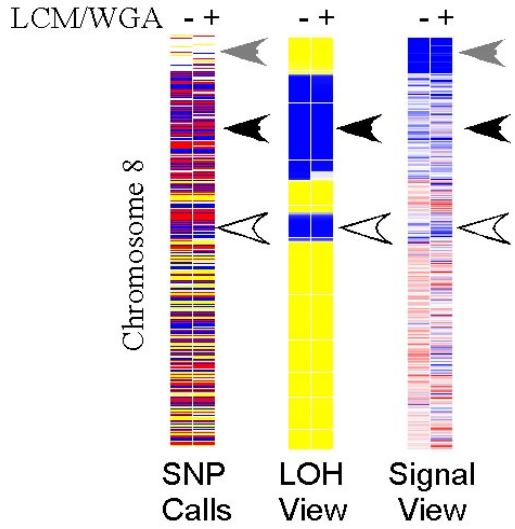


Figure 2: High-quality SNP array data generated after laser capture microdissection and whole genome amplification. Top: Data generated from a highly pure tumor after macrodissection is displayed alongside data from the same sample after LCM and WGA. Genotype calls are displayed in the left panel (yellow – AB (het), red – AA, blue – BB, white – No Call), LOH inferences in the middle panel (blue – LOH, yellow – retention), and signal intensities in the right panel (red – high, white – normal, blue – low). A homozygous deletion on chr. 8 (gray arrow) and regions of LOH (black and white arrows) are indicated.

Likewise, the ability to identify copy number aberrations is preserved (Figures 2,3). The main concern here is due to the potential uneven nature of amplification by WGA. In fact, we know that certain regions of the genome are better represented in WGA product than others, with up to 6-fold variations between different regions (Dean et al., 2002). As long as these biases are consistent, however, normalization against control samples that have undergone whole genome amplification under similar conditions will correct for them, leaving only the underlying signal intensity changes reflecting copy number aberrations inherent in the sample. To test how consistent these biases are along a range of WGA conditions, DNA obtained from laser capture microdissected benign prostate tissue was amplified under varying conditions.

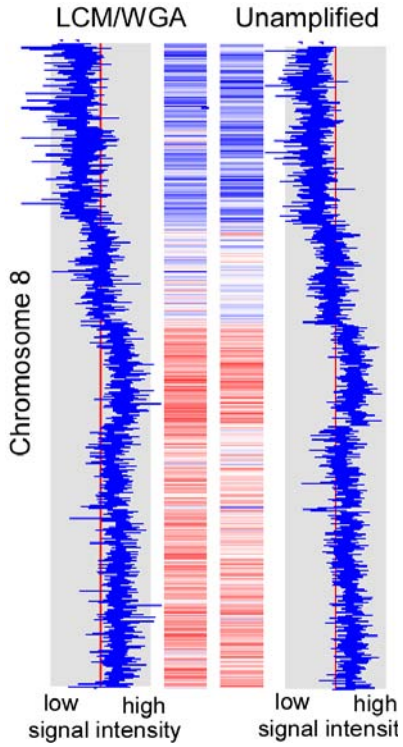


Figure 3: Preservation of signal intensity variations after laser capture microdissection and whole genome amplification. Normalized signal intensities for a second highly pure tumor are displayed (as in Figure 5, right panel), along with graphs of those signal intensities on either side. Both amplifications and deletions appear highly reproducible after laser capture microdissection and whole genome amplification.

Specifically, as the amount of WGA template was increased from 4 ng to 64, the SNP array signal intensities became more similar to unamplfiied controls (Table 1). However, signal intensities from LCM/WGA were highly consistent, when compared against DNA undergoing LCM/WGA under similar conditions (mean variance 0.11, same as unamplified controls;

Table 1). Moreover, this consistency remained even when DNA from LCM was whole genome amplified on different days (variance = 0.12), or using different lots of polymerase, dNTP, and random primers (variance = 0.10). Finally, the signal intensities from LCM/WGA appear robust to 2-fold changes in amount of DNA template (variance = 0.13).

Table 1. Mean variance between normalized SNP array signal intensities obtained from WGA product using various amounts of template DNA, versus signal intensities from unamplified DNA. Template DNA used for WGA was produced from LCM of benign prostate tissue.

Template amount	4 ng	8 ng	16 ng	32 ng	64 ng	Unamplified*
Variance from unamplified DNA	0.30	0.26	0.24	0.23	0.23	0.11
Variance from similarly prepared DNA	-	0.13	-	0.09	-	0.11

*Variance between normalized signal intensities from repeat SNP arrays using the same unamplified DNA

b. Collection of primary and metastatic tumors for microdissection

Through the Gelb Center at the Dana-Farber Cancer Institute, we have IRB- approved access to several hundred fresh frozen primary prostate cancers, along with uninvolved seminal vesicles. To date, Gleason 3+3 (n=19), 3+4 (n=12), 4+3 (n=7), Gleason 4+4 (n=11), and Gleason 4+5 (n=1) have been selected for microdissection and analysis.

Metastatic tumors are being obtained from several sources. Through our collaborator, Dr Mark Rubin, we have obtained hormone-naïve lymph node metastases from 6 individuals, with sufficient tumor to obtain adequate amounts of DNA. Dr Rubin has also coordinated the gathering of metastatic tissue from 18 individuals, obtained through the rapid autopsy program at the University of Michigan. Through Dr Steven Balk, we have obtained multiple bone marrow biopsy specimens, and found that 3 contain sufficient tumor for microdissection. In addition, Dr Mary-Ellen Taplin has provided a further 8 bone marrow biopsy samples. Recently, we have initiated a collaboration with Dr Lawrence True and his colleagues at the University of Washington, and through them have obtained metastatic tissue from 20 individuals, garnered through their rapid autopsy program, along with primary prostate tissue from the majority of these individuals. Histologic characterization and microdissection of all of these samples is underway.

c. Production of high-quality DNA from whole genome amplification of DNA obtained from laser-capture microdissected tumors and germline tissue

With tumors collected and optimal conditions for LCM and WGA determined, we initiated data collection by performing LCM on primary tumors along with germline tissue from paired seminal vesicles. Unfortunately, we soon came to find that histology affects the quality of DNA obtained after LCM and WGA. For instance, in one experiment we performed LCM and WGA on 3 primary prostate cancers along with paired uninvolved seminal vesicles. In each case, 32 ng of DNA was used from the laser captured cells as template for WGA, and 250 ng of WGA product was used for restriction digest, amplification, and hybridization to 50K Xba arrays. WGA was performed using the same reagents and at the same time, for all samples. However, genotyping call rates were excellent for the germline DNA obtained from the seminal vesicles, ranging from 95-98%, and were low for the tumors, ranging from 85-90%. Moreover, signal intensity profiles were much noisier for the tumor DNA (data not shown), precluding high-resolution copy number analysis.

As both the tumor and normal seminal vesicle were resected from the patient simultaneously, the cause of the difference in DNA quality between them is likely due to the differing histology between the tumor and normal tissue and resulting differences in the laser capture process itself. Namely, small nests

of cancer cells have to be captured from the tumor, whereas large regions of normal tissue could be captured. The result is that the captured tumor cells, on average, lie closer to the line cut by the laser. We are testing the possibility that the laser itself is damaging DNA. In the meantime, we have purchased an Arcturus Veritas laser capture microdissection machine, and have begun experiments with capturing cells using an infrared rather than ultraviolet laser, which is likely to reduce radiation damage.

In case this approach does not work, we are also collecting highly pure primary tumors, for which we can microdissect sufficient numbers of tumor cells so as to not require whole genome amplification. Prior experiments have shown even in cases where WGA DNA performs poorly on the SNP arrays due to slight degradation of the template DNA, the unamplified template DNA will still perform well (data not shown).

2. Specific aim #2: To generate genome-wide high-resolution maps of LOH and copy-number alterations using SNP arrays containing probes for 100,000 markers.

While optimizing the LCM methods, unamplified DNA was used to produce 100K SNP array data from 45 prostate cancers, including 11 primary tumors, 4 hormone-naïve lymph node metastases, 7 hormone-refractory metastases, 7 cell lines, and 16 xenografts.

a. Generation of LOH maps

In the application for this award, we described a method we had developed to identify regions of LOH without the use of paired normal DNA. Although we are obtaining paired normal DNA for all primary and metastatic tumors in this study, we also have SNP array data from prostate cancer model systems for which paired normal DNA is unavailable. The method used to determine LOH without paired normal DNA was originally developed using data from SNP arrays probing 10,000 loci throughout the genome. When applying the method to 100K SNP array data, we found that the haplotype structure of the genome reduced the specificity of the method, and improved the method to take this haplotype structure into account (Beroukhim et al, in review; Appendix I).

Those regions that most frequently undergo LOH are most likely to influence tumor survival, through the presence of tumor suppressor genes (TSGs). In fact, our analysis of the frequency of LOH in 34 prostate cancer samples revealed that several tumor suppressor genes lay at regions of peak LOH frequency, compared to the rest of the genome (Figure 7). Here, we use as a measure of LOH frequency the “LOH probability mean”, which refers to the mean probability with which LOH occurred at each SNP locus, averaged across all 34 samples.

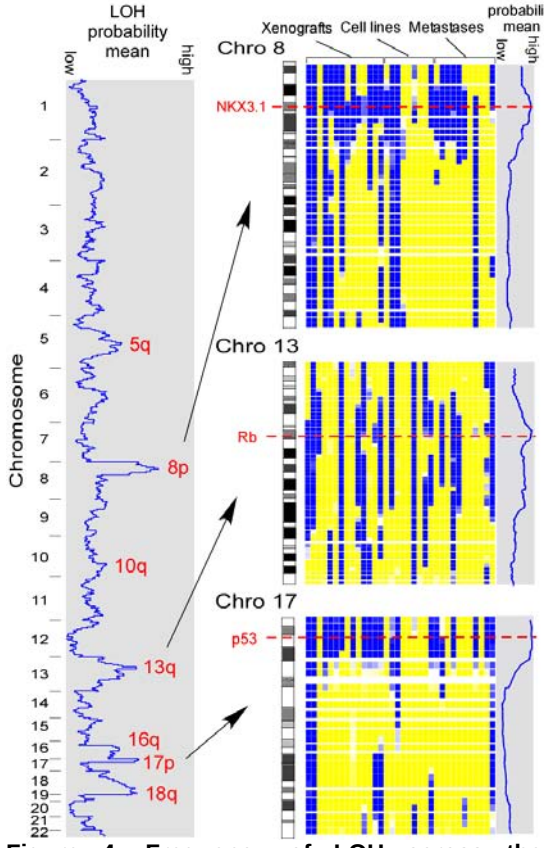


Figure 4: Frequency of LOH across the prostate cancer genome. The mean LOH probability across 34 prostate cancer samples is plotted along the left for all chromosomes. Peak regions of LOH are noted, and data from chromosomes 8, 13, and 17 are highlighted on the right. These data are displayed as in Figure 2. Note that in this view, SNPs are visualized proportional to physical distance along the chromosome and most SNPs are not projected due to proximity to their neighbors. The red dotted lines indicate the approximate chromosomal positions of putative TSGs.

b. Generation of copy-number maps

For the generation of copy-number estimates, both systematic and random errors in signal intensity data have to be minimized. We found that the main source of systematic error is *batch effect*, whereby a batch of samples that simultaneously undergo DNA digestion, amplification, labeling, and hybridization to arrays, will have similar signal intensity alterations (high or

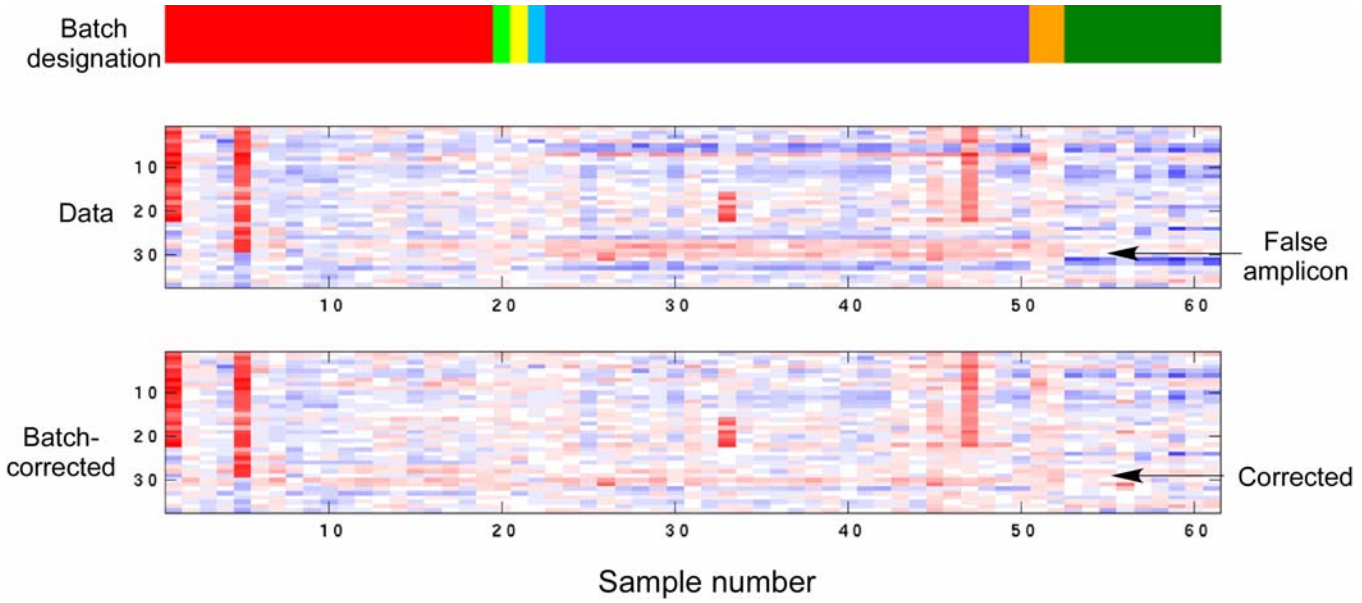


Figure 5: Batch effect and correction. Signal intensity data are displayed for 61 samples, each as a column. These samples were run in 7 batches, designated by different colors in the top panel. The normalized intensities for a set of 40 consecutive SNPs are displayed (as in Figure 2) in the middle panel. In the batch designated in blue (top panel), several adjacent SNPs appear to have consistently high signal intensities, giving the false appearance of a recurrent amplicon. Data corrected for this batch effect are displayed in the bottom panel.

low signal intensity) compared to samples processed at other times. This, in turn, leads to the appearance of amplicons and deletions restricted to that batch (Figure 5). Strict control of experimental conditions and normalization against reference samples from the same batch can minimize, but tend not to eliminate, these batch effects. In turn, these batch effects can lead to the identification of spurious regions of recurrent amplification and deletion.

We posited that the identifying characteristic of an alteration due to batch effect is that the alteration consistently occurs within one batch, and consistently does not occur within other batches of similar samples. Therefore, for each batch containing at least 5 samples, we identified the distribution of signal intensities for each SNP, and compared this using a T-test with the distribution of signal intensities in all other batches. When the p value was less than 0.001, we considered that the SNP had undergone a systematic alteration due to batch effect, and subtracted a constant amount from the signal intensities at that SNP for all samples in the batch, so that the mean signal in that batch equaled the mean signal in all other batches (Figure 5).

Whereas systematic errors can lead to the identification of spurious regions of amplification and deletion, random errors tend to reduce our sensitivity to identifying regions of real importance. Most importantly, the error in the signal intensity measured at a given SNP can lead to that SNP being spuriously identified as amplified or deleted, leading to downstream errors in estimates of the frequency of lesions at that SNP locus. A variety of smoothers, developed for CGH data, reduce noise levels at each locus by involving information from neighboring loci (Lai et al., 2005). We have found GLAD (Hupe et al., 2004), which identifies segments with a constant copy number and averages the signal intensities across all loci in each segment, provides the most accurate results in a reasonable amount of computational time (data not shown). Several alternative software packages (dChipSNP, CNAT, CNAG, GIM) (Bignell et al., 2004; Ishikawa et al., 2005; Nannya et al., 2005; Zhao et al., 2004) also exist to convert probe-level data into overall SNP-specific signal intensities. Preliminary results seem to point to CNAG as producing the most optimal signal-to-noise ratios (data not shown).

c. Identification of significant copy-number aberrations

We posited that information as to the importance of a region in sustaining cancer lay not only in the frequency with which that region undergoes lesions, but also the also in the amplitude of the lesions that occur. Therefore, we designed scores for amplification and deletion that included both sources of information. Namely, for each SNP locus we calculate:

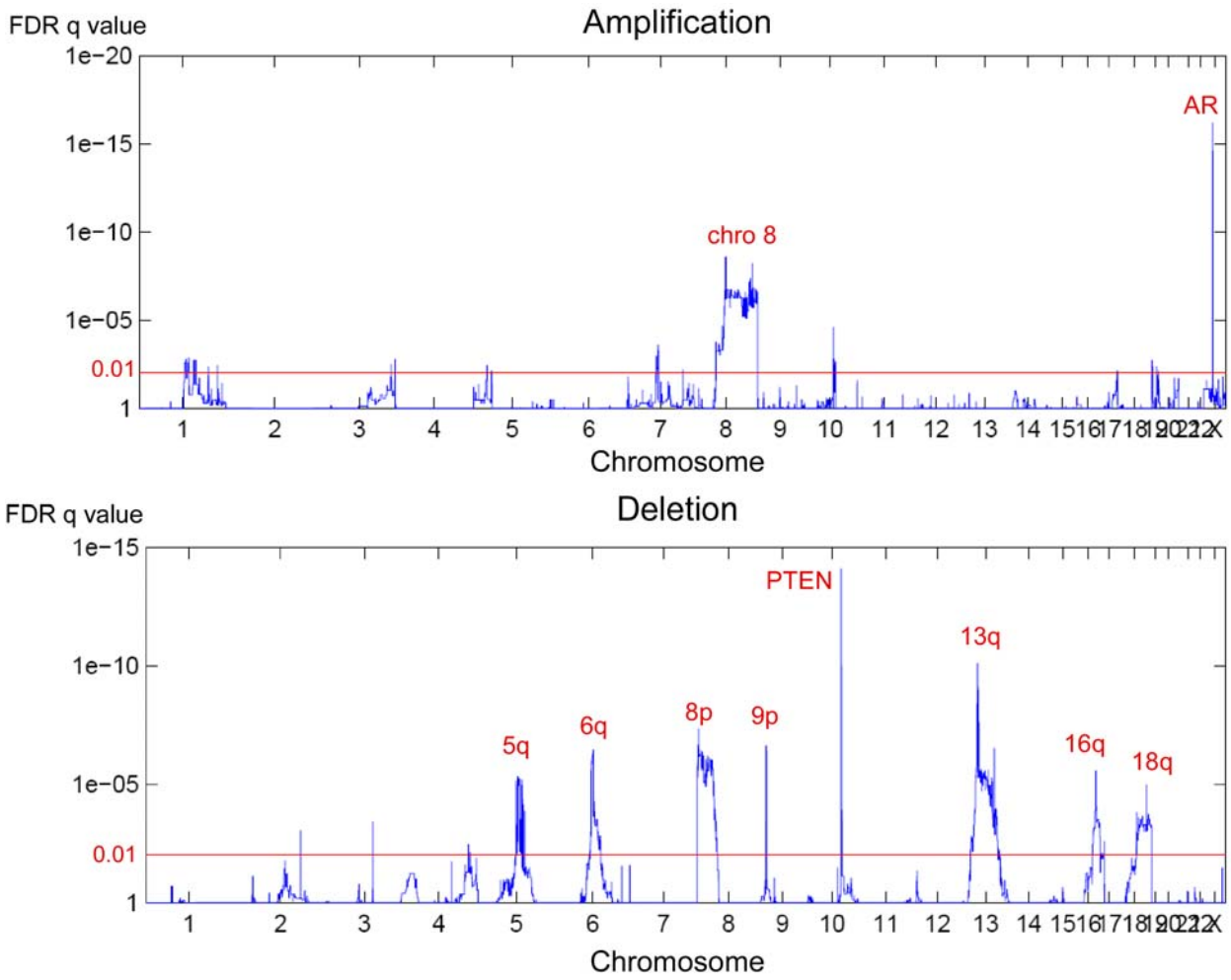


Figure 6: Regions of significant amplification and deletion in a set of 45 prostate cancers. FDR-corrected q values (log scale, left), associated with Amp and Del scores across the genome, are displayed. Regions with a q value less than 0.01 (red line) were considered significantly altered. Among amplifications, the most significant region overlaps the androgen receptor (labeled as AR). Among deletions, the region containing PTEN scored as most significant.

$$\text{Amp} = f_{amp} \times \log_2(\hat{S}_{amp}), \text{ and}$$
$$\text{Del} = f_{del} \times -\log_2(\hat{S}_{del}) \quad (1)$$

where f_{amp} and f_{del} represent the frequency of amplification and deletion, respectively, and $\hat{S}_{amp}$ and $\hat{S}_{del}$ represent the average normalized signal intensity of samples with amplifications and deletions.

The significance of each particular Amp or Del score is then determined by comparing it to similar scores determined from all permutations of the data, allowing the calculation of p values and, to correct for multiple hypothesis testing, False Discovery Rate (FDR) q values (Benjamini and Hochberg, 1995).

When applied to our data from 45 prostate cancers, we obtained the Amp and Del scores displayed in Figure 9. Regions of amplification and deletion having q values less than 0.01 (i.e. having less than a 1% probability of occurring by chance alone) were designated as significant. The region surrounding the androgen receptor is the most significantly amplified, whereas the region surrounding PTEN is the most significantly deleted. Interestingly, PTEN deletions (including homozygous deletions) are highly significant (Figure 6), but LOH of broader areas of 10q is not as common (Figure 5). Conversely, LOH of 17p is very common (Figure 5), but deletions of this region are less significant (Figure 6)—due in part to a high prevalence of copy neutral LOH at this locus. These results underline the importance of obtaining independent maps of deletional and LOH events.

Multiple other regions of significant amplification and deletion are seen, although the targets of most of these regions are not known. By adding higher-resolution data from larger numbers of tumors, including more primary and metastatic tissue samples, the locations of the most significant regions will become increasingly refined.

3. Specific aim #3: To identify and validate candidate somatic genetic alterations differing in prevalence between localized and metastatic cancers, and develop markers for clinical association studies.

Table 2. Prevalence of selected lesions in prostate cancer samples of varying stages.

Lesion	Primaries (n=11)	Mets (n=11)	p-value
9p loss	0	8	0.001
PTEN loss	0	8	0.001
18q loss	1	8	0.008
13q loss	5	8	0.39
8p loss	9	9	1
16q23 loss	9	9	1
17p loss	8	8	1

	hormone-naïve* (n=15)	hormone-refractory (n=7)	p-value
AR amp	1	6	0.0006
21q22 loss	3	6	0.007

* includes primaries

The two groups differ

The two groups are similar

The analysis in the prior section not only identified the significant regions of amplification and deletion in these 45 prostate cancer samples; to do so, each sample had to be independently characterized as to the lesions it harbors. Thus, we can immediately identify correlations between the presence of any two or more lesions, as well as correlations between the presence of a set of lesions and phenotype. Although the sample set is small, we have begun this analysis in the setting of prostate cancer progression. Namely, we can begin to distinguish genetic lesions that are equally prevalent between primary prostate cancers and metastases, from those that are more common among hormone-naïve or hormone-refractory prostate cancer (Figure 7, Table 2). For example, loss of 9p and PTEN were only seen in metastases, whereas loss of 8p, 16q, and 17p were equally prevalent among primaries and metastases. As expected, androgen receptor (AR) amplification tends to occur in hormone-refractory samples, with only one hormone-naïve lymph node metastasis having a low-level amplification of AR. However, while high-level amplifications of AR have been reported, low-level gene duplications are seen in 2 of the 7 samples with AR amplification. This accords well with the finding that only modest overexpression of AR is required for hormone resistance to be achieved (Chen et al., 2004). Interestingly, 21q22 loss also appears more prevalent among the hormone-refractory than hormone-naïve samples for which we have data.

Despite the small numbers of samples in this preliminary dataset, the p values for several lesions have already attained statistical significance, although they have not been corrected for multiple hypothesis testing.

4. Key Research Accomplishments

- Developed method for determination of LOH without paired normals that takes into account haplotype structure
- Developed methods for reducing signal-intensity errors, including systematic errors due to batch effects
- Developed methods for identifying significant regions of copy-number aberration
- Correlated several regions with progressive cancer

5. Reportable Outcomes

- Manuscript in review: **Beroukhim R**, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, Descazeaud A, Rubin MA, Meyerson M, Wong WH, Sellers WR, and Li C, “Inferring Loss-of-Heterozygosity From Tumors Without Paired Normals Using High-Density SNP Arrays”.
- Publication: Mellinghoff IK, Wang MY, Vivanco I, Haas-Kogan DA, Zhu S, Dia EQ, Lu KV, Yoshimoto K, Huang JH, Chute DJ, Riggs BL, Horvath S, Liau LM, Cavenee WK, Rao PN, **Beroukhim R**, Peck

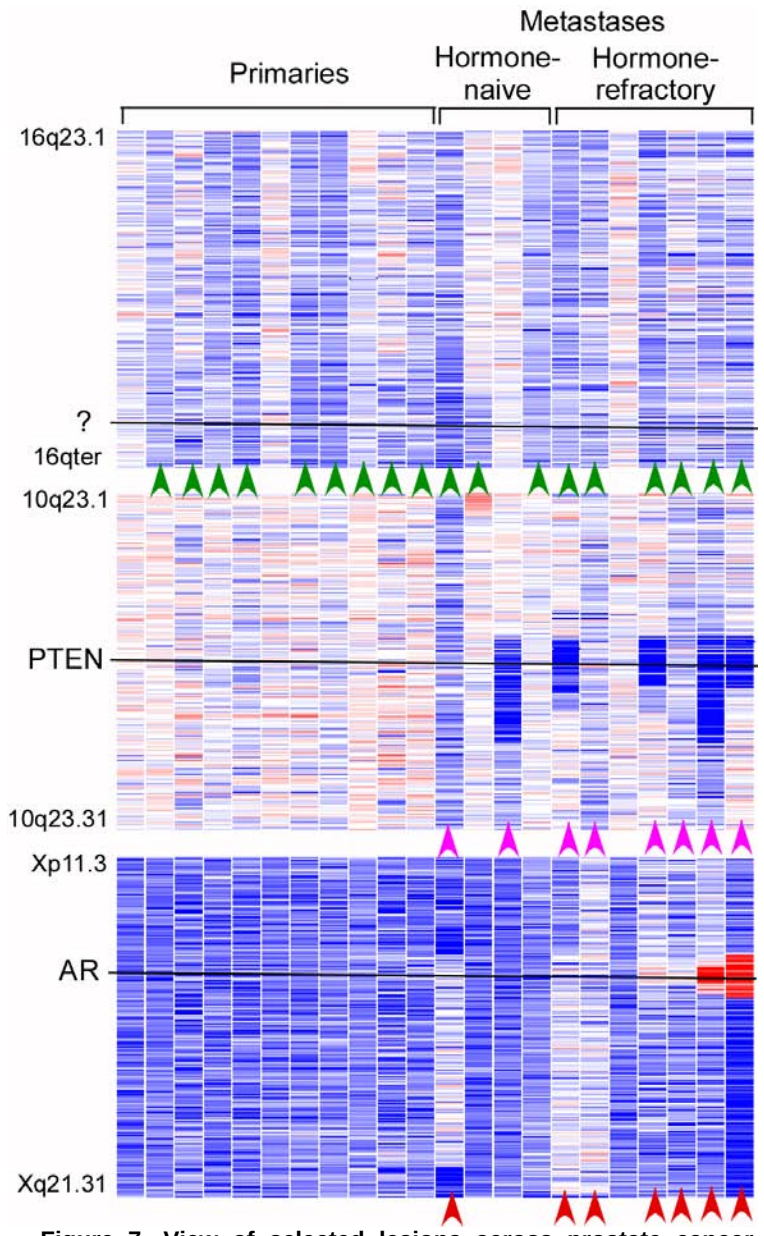


Figure 7: View of selected lesions across prostate cancer samples of varying stages. Signal intensity data are displayed (as in Figure 2) for regions of chromosomes 16q, 10q, and X.
Upper Panel: Hemizygous loss (with LOH; seen here in blue) is a frequent event in primary and metastatic prostate cancers. Green arrows label samples observed to have loss. Although ATBF1 has been suggested as the targeted tumor suppressor gene(Sun et al., 2005), the most frequent focus of deletion in our samples is located more towards the q terminus (question mark).
Middle Panel: Homozygous and hemizygous losses of PTEN are seen in 2 hormone-naïve and 6 hormone-refractory metastases (purple arrows), but not in any of the primary cancers examined to date.
Bottom Panel: High-level gains of the androgen receptor are seen in 4 hormone-refractory metastases; another 2 have low-level amplifications, as does one hormone-naïve metastasis (red arrows). None of the primary cancers examined to date appear to have any level of androgen receptor amplification.

of K, in G, Li TC, Lee JC, Sellers WR, Stokoe D, Prados M, Cloughesy TF, Sawyers CL, Mischel PS, "Molecular determinants of the response glioblastomas to EGFR kinase inhibitors", NEJM. 2005;353:2012-24.

- Publication: Garraway LA, Weir BA, Zhao X, Widlund H, **Beroukhim R**, Berger A, Rimm D, Rubin MA, Fisher DE, Meyerson ML, Sellers WR, "'Lineage Addiction' in Human Cancer: Lessons from Integrated Genomics", Cold Spring Harb Symp Quant Biol. 2005;70:1-10.
- Publication: Koochekpour S, Zhuang YJ, **Beroukhim R**, Hsieh CL, Hofer MD, Zhau HE, Hiraiwa M, Pattan DY, Ware JL, Luftig RB, Sandhoff Sawyers CL, Pienta KJ, Rubin MA, Vessella RL, Sellers WR, Sartor O, "Amplification and overexpression of prosaposin prostate cancer", Genes Chromosomes Cancer. 2005; 44:351-64.
- Publication: Garraway LA, Widlund HR, Rubin MA, Getz Berger AJ, Ramaswamy S, **Beroukhim R**, Milner DA, Granter SR, Du J, Lee C, Wagner SN, Li C, Golub TR, Rimm DL, Meyerson ML, Fisher DE, Sellers WR, "Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma", Nature. 2005; 436:117-22
- Publication: Zhao X, Weir BA, LaFramboise T, Lin M, **Beroukhim R**, Garraway L, Beheshti J, Lee JC, Naoki K, Richards WG, Sugarbaker D, Chen F, Rubin MA, Janne PA, Girard L, Minna J, Christiani D, C, Sellers WR, Meyerson M, "Homozygous deletions and chromosome amplifications in human lung carcinomas revealed by single nucleotide polymorphism array analysis", Cancer Res. 2005; 65:5561-70.

6. Conclusions

In the 11 months since the beginning of the award, significant progress has been made in all 3 specific aims of this grant, including determination of LOH and copy number maps using 100K SNP array data from 45 prostate tumors, and use of these maps to identify chromosomal aberrations that appear to be playing a significant role in prostate cancer. Some of these regions appear to correlate with prostate cancer progression. However, unanticipated difficulties have also arisen with respect to laser capture microdissection of primary prostate tissue, which appears to provide lower-quality DNA than laser capture microdissection of paired normal tissue. The reasons for these difficulties are being investigated, and modifications to the LCM protocol, that will avoid these difficulties, are possible.

7. References

Allinen, M., Beroukhim, R., Cai, L., Brennan, C., Lahti-Domenici, J., Huang, H., Porter, D., Hu, M., Chin, L., and Richardson, A. (2004). Molecular characterization of the tumor microenvironment in breast cancer. *6*, 17.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J Roy Stat Soc, Ser B *57*, 289-300.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R.*, et al.* (2004). High-Resolution Analysis of DNA Copy Number Using Oligonucleotide Microarrays. Genome Res *14*, 287-295.

Chen, C. D., Welsbie, D. S., Tran, C., Baek, S. H., Chen, R., Vessella, R., Rosenfeld, M. G., and Sawyers, C. L. (2004). Molecular determinants of resistance to antiandrogen therapy. Nature Medicine *10*, 33.

Craig, D. W., and Stephan, D. A. (2005). Applications of whole-genome high-density SNP genotyping. Expert Review of Molecular Diagnostics *5*, 159-170.

Cutler, D. J., Zwick, M. E., Carrasquillo, M. M., Yohn, C. T., Tobin, K. P., Kashuk, C., Mathews, D. J., Shah, N. A., Eichler, E. E., Warrington, J. A., and Chakravarti, A. (2001). High-Throughput Variation Detection and Genotyping Using Microarrays. Genome Res *11*, 1913-1925.

Dean, F. B., Hosono, S., Fang, L., Wu, X., Faruqi, A. F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J.*, et al.* (2002). Comprehensive human genome amplification using multiple displacement amplification. PNAS *99*, 5261-5266.

Dumur, C. I., Dechsukhum, C., Ware, J. L., Cofield, S. S., Best, A. l. M., Wilkinson, D. S., Garrett, C. T., and Ferreira-Gonzalez, A. (2003). Genome-wide detection of LOH in prostate cancer using human SNP microarray technology. *81*, 260.

Garraway, L. A., Widlund, H. R., Rubin, M. A., Getz, G., Berger, A. J., Ramaswamy, S., Beroukhim, R., Milner, D. A., Granter, S. R., Du, J.*, et al.* (2005). Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature *436*, 117.

Hoque, M. O., Lee, C.-C. R., Cairns, P., Schoenberg, M., and Sidransky, D. (2003). Genome-Wide Genetic Characterization of Bladder Cancer: A Comparison of High-Density Single-Nucleotide Polymorphism Arrays and PCR-based Microsatellite Analysis. Cancer Res *63*, 2216-2222.

Huang, J., Wei, W., Zhang, J., Liu, G., Bignell, G. R., Stratton, M. R., Futreal, P. A., Wooster, R., Jones, K. W., and Shapero, M. H. (2004). Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Human Genomics *1*, 287-299.

Hughes, S., Arneson, N., Done, S., and Squire, J. (2005). The use of whole genome amplification in the study of human disease. *88*, 173.

Hupe, P., Stransky, N., Thiery, J.-P., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. Bioinformatics *20*, 3413-3422.
Irving, J. A. E., Bloodworth, L., Bown, N. P., Case, M. C., Hogarth, L. A., and Hall, A. G. (2005). Loss of Heterozygosity in Childhood Acute Lymphoblastic Leukemia Detected by Genome-Wide Microarray Single Nucleotide Polymorphism Analysis. Cancer Res *65*, 3053-3058.
Ishikawa, S., Komura, D., Tsuji, S., Nishimura, K., Yamamoto, S., Panda, B., Huang, J., Fukayama, M., Jones, K. W., and Aburatani, H. (2005). Allelic dosage analysis with genotyping microarrays. *333*, 1309.
Jacqueline A. Langdon, J. M. L. D. K. S. S. D. E. P. N. B. R. G. G. D. W. E. S. C. C. (2005). Combined genome-wide allelotyping and copy number analysis identify frequent genetic losses without copy number reduction in medulloblastoma. Genes, Chromosomes and Cancer *9999*, NA.
Janne, P. A., Li, C., Zhao, X., Girard, L., Chen, T.-H., Minna, J., Christiani, D. C., Johnson, B. E., and Meyerson, M. (2004). High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene *23*, 2716-2726.
Lai, W. R., Johnson, M. D., Kucherlapati, R., and Park, P. J. (2005). Comparative analysis of algorithms for identifying amplifications and deletions in array CGH data. Bioinformatics, bti611.
Lieberfarb, M. E., Lin, M., Lechpammer, M., Li, C., Tanenbaum, D. M., Febbo, P. G., Wright, R. L., Shim, J., Kantoff, P. W., Loda, M.*, et al.* (2003). Genome-wide Loss of Heterozygosity Analysis from Laser Capture Microdissected Prostate Cancer Using Single Nucleotide Polymorphic Allele (SNP) Arrays and a Novel Bioinformatics Platform dChipSNP. Cancer Res *63*, 4781-4785.
Lindblad-Toh, K., Tanenbaum, D. M., Daly, M. J., Winchester, E., Lui, W.-O., Villapakkam, A., Stanton, S. E., Larsson, C., Hudson, T. J., Johnson, B. E.*, et al.* (2000). Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. *18*, 1001.
Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H.*, et al.* (2004). Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods *1*, 109.
Matsuzaki, H., Loi, H., Dong, S., Tsai, Y.-Y., Fang, J., Law, J., Di, X., Liu, W.-M., Yang, G., Liu, G.*, et al.* (2004). Parallel Genotyping of Over 10,000 SNPs Using a One-Primer Assay on a High-Density Oligonucleotide Array. Genome Res *14*, 414-425.
Mei, R., Galipeau, P. C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R. K., Chee, M. S., Reid, B. J., and Lockhart, D. J. (2000). Genome-wide Detection of Allelic Imbalance Using Human SNPs and High-density DNA Arrays. Genome Res *10*, 1126-1137.
Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., Kurokawa, M., Chiba, S., Bailey, D. K., Kennedy, G. C., and Ogawa, S. (2005). A Robust Algorithm for Copy Number Detection Using High-Density Oligonucleotide Single Nucleotide Polymorphism Genotyping Arrays. Cancer Res *65*, 6071-6079.
Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X., Richter, D. J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D.*, et al.* (2004). Genome coverage and sequence fidelity of {phi}29 polymerase-based multiple strand displacement whole genome amplification. Nucl Acids Res *32*, e71-.
Primdahl, H., Wikman, F. P., von der Maase, H., Zhou, X.-g., Wolf, H., and Orntoft, T. F. (2002). Allelic Imbalances in Human Bladder Cancer: Genome-Wide Detection With High-Density Single-Nucleotide Polymorphism Arrays. J Natl Cancer Inst *94*, 216-223.
Rubin, M. A., Varambally, S., Beroukhim, R., Tomlins, S. A., Rhodes, D. R., Paris, P. L., Hofer, M. D., Storz-Schweizer, M., Kuefer, R., Fletcher, J. A.*, et al.* (2004). Overexpression, Amplification, and Androgen Regulation of TPD52 in Prostate Cancer. Cancer Res *64*, 3814-3822.

Sachidanandam, R., Weissman, D., Schmidt, S., Kakol, J., Stein, L., Marth, G., Sherry, S., Mullikin, J., Mortimore, B., Willey, D.*, et al.* (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *409*, 928.
Schubert, E. L., Hsu, L., Cousens, L. A., Glogovac, J., Self, S., Reid, B. J., Rabinovitch, P. S., and Porter, P. L. (2002). Single Nucleotide Polymorphism Array Analysis of Flow-Sorted Epithelial Cells from Frozen Versus Fixed Tissues for Whole Genome Analysis of Allelic Loss in Breast Cancer. Am J Pathol *160*, 73-79.
Sun, X., Frierson, H. F., Chen, C., Li, C., Ran, Q., Otto, K. B., Cantarel, B. M., Vessella, R. L., Gao, A. C., Petros, J.*, et al.* (2005). Frequent somatic mutations of the transcription factor ATBF1 in human prostate cancer. *37*, 407.
Wang, Z. C., Lin, M., Wei, L.-J., Li, C., Miron, A., Lodeiro, G., Harris, L., Ramaswamy, S., Tanenbaum, D. M., Meyerson, M.*, et al.* (2004). Loss of Heterozygosity and Its Correlation with Expression Profiles in Subclasses of Invasive Breast Cancers. Cancer Res *64*, 64-71.
Zhao, X., Li, C., Paez, J. G., Chin, K., Janne, P. A., Chen, T.-H., Girard, L., Minna, J., Christiani, D., Leo, C.*, et al.* (2004). An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays. Cancer Res *64*, 3060-3071.
Zhao, X., Weir, B. A., LaFramboise, T., Lin, M., Beroukhim, R., Garraway, L., Beheshti, J., Lee, J. C., Naoki, K., Richards, W. G.*, et al.* (2005). Homozygous Deletions and Chromosome Amplifications in Human Lung Carcinomas Revealed by Single Nucleotide Polymorphism Array Analysis. Cancer Res *65*, 5561-5570.
Zhou, X., Mok, S. C., Chen, Z., Li, Y., and Wong, D. T. W. (2004). Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. Human Genetics *115*, 327.

Inferring Loss-of-Heterozygosity From Tumors Without Paired Normals Using High-Density SNP Arrays

Rameen Beroukhim[†],[1,2,3] Ming Lin[†],[1,4] Yuhyun Park,[1,4] Ke Hao[4], Xiaojun Zhao,[1,3] Levi A. Garraway,[1,2,3] Edward A. Fox,[1] Ephraim P. Hochberg,[1,2,3,5] Ingo K. Mellinghoff,[6] Matthias D. Hofer,[2,3] Aurelien Descazeaud,[2,3] Mark A. Rubin,[2,3] Matthew Meyerson,[1,3,7] Wing Hung Wong,[8] William R. Sellers[*1,2,3,7] and Cheng Li[*1,4]

[1] Departments of Biostatistics and Computational Biology and Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, 02115 [2] Departments of Medicine, Pathology, and Radiation Oncology, Brigham and Women's Hospital, Boston, MA, 02115 [3] Departments of Medicine and Pathology, Harvard Medical School, Boston, MA, 02115 [4] Department of Biostatistics, Harvard School of Public Health, Boston, MA, 02115 [5] Department of Medicine, Massachusetts General Hospital [6]Departments of Medicine and Medical Pharmacology, David Geffen School of Medicine, University of California, Los Angeles, CA, 90095 [7] Broad Institute of Harvard and MIT, 320 Charles Street, Cambridge, MA, 02141 [8] Department of Statistics, Stanford University, Stanford, CA, 94305

† Contributed equally.

* Correspondence: William R. Sellers, Department of Medical Oncology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, Phone: 617-632-5261, Fax: 617-632-3460, William_Sellers@dfci.harvard.edu

Cheng Li, Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA, 02115, Phone: 617-632-3498, Fax: 617-632-5444, cli@hsph.harvard.edu

Running title: Tumor-only LOH inference of SNP array data

Keywords: loss of heterozygosity, single nucleotide polymorphism, oligonucleotide microarray, allelic imbalance

ABSTRACT

Background: Loss of heterozygosity (LOH) of chromosomal regions bearing tumor suppressors is a key event in the evolution of epithelial and mesenchymal tumors. Identification of these regions usually relies on genotyping tumor and counterpart normal DNA and noting regions where heterozygous alleles in the normal become homozygous in the tumor. However, paired normal samples for tumors and cell lines are often not available. With the advent of oligonucleotide arrays that simultaneously assay thousands of single nucleotide polymorphism (SNP) markers, genotyping can now be done at high enough resolution to allow identification of LOH events by the absence of heterozygous loci, without comparison to normal controls.

Methodology/Principle Findings: We describe a Hidden Markov Model based method to identify LOH from unpaired tumor samples, taking into account SNP intermarker distances, SNP-specific heterozygosity rates, and the haplotype structure of the human genome. When we applied the method to data genotyped on 100K arrays, we correctly identified 99% of SNP markers as either retention or loss. We also correctly identified 81% of the regions of LOH, including 98% of regions greater than 3 Mb. By integrating copy number analysis into the method, we were able to distinguish LOH from allelic imbalance. Application of this method to data from a set of prostate samples without paired normals identified known regions of prevalent LOH.

Conclusions/Significance: We have developed a method for analyzing high-density oligonucleotide SNP array data to accurately identify of regions of LOH and retention in tumors without the need for paired normal samples.

INTRODUCTION

Loss of heterozygosity (LOH) refers to change from a state of heterozygosity in a normal genome to a homozygous state in a paired tumor genome. LOH is most often regarded as a mechanism for disabling tumor suppressor genes (TSGs) during the course of oncogenesis [1,2]. Although LOH is often thought to result from copy-loss events such as hemizygous deletions, a large proportion of LOH results from copy-neutral events such as chromosomal duplications [3,4]. Analyzing LOH data across multiple tumor samples can point to loci harboring TSGs or identify subtypes of tumors with different somatic genetic profiles [5,6].

Single nucleotide polymorphisms (SNPs) are the most common genetic variation in the human genome and can be used to search for germline genetic contributions to disease. To that end, oligonucleotide SNP arrays have been developed to simultaneously genotype thousands of SNP markers across the human genome [7-9]. The density, distribution, and allele specificity of SNPs makes them attractive for high-resolution analyses of LOH and copy number alterations in cancer genomes [3,6,10-15].

Traditionally, LOH analyses require the comparison of the genotypes of the tumor and its normal germline counterpart. However, for cell lines, xenograft, leukemia and archival samples, paired normal DNA is often unavailable. Current generations of SNP arrays provide high enough marker density to make it feasible to identify regions of LOH by the absence of heterozygous loci (which we call inferred LOH), rather than by comparison to the paired normal. For example, the homozygosity mapping of deletions (HOMOD) method was developed to use highly polymorphic microsatellite markers to identify regions of hemizygous deletion in unpaired tumor

cell lines [16], and a simple method of inferring LOH using the product of the probability of homozygosity in neighboring SNPs was able to identify 80% of LOH in 10K SNP array data from one sample [3,17]. SNP markers are less polymorphic than microsatellite markers, however, and the haplotype structure may render closely located SNP dependent in their genotype calls. We hypothesized that a method that infers LOH with high accuracy would have to account for not only the varied heterozygosity rates of SNP markers, but also their varied intermarker distances, as well as genotyping errors and the interdependence of SNP alleles based on the haplotype structure of the genome.

We approached this problem by developing a Hidden Markov Model (HMM) to infer LOH. HMMs are appropriate for inferring the unobserved underlying states that give rise to an observed chain of data, using multiple sources of information. They have been used to model biological data in diverse applications such as sequence analysis [18-20], linkage studies [21,22] and array comparative genomic hybridization [15,23]. SNP genotypes along a chromosome are chain-like and thus suitable for HMM analysis. The model we developed incorporates SNP intermarker distances, SNP-specific heterozygosity rates, and genotyping error rate. We show that it accurately identifies regions of LOH in unpaired tumors. We find that when genotyping data is obtained at very high densities (100,000 markers across the genome), regions of false LOH are identified unless the haplotype block structure of the genome is taken into account and used to modify the HMM accordingly. Integrating copy number analysis allows the distinction of LOH from allelic imbalance. Application of this method to data from prostate cell lines, xenografts, and metastases lacking paired normals identifies known regions of prevalent LOH, containing known and putative TSGs.

METHODS

Tumor samples and paired normals. We used data from Early Access 10K, Mapping 10K and 100K SNP arrays (Affymetrix) (referred to as 10K, 11K and 100K arrays respectively) interrogating, respectively, 10,044; 11,555; and 116,204 SNP loci on all chromosomes except Y, with an average intermarker distance of 210 kb (11K array) and 23.6 kb (100K array) and average heterozygosity rate of 0.38 (11K array) and 0.27 (100K array) [7-9]. 10K array data from paired tumor/normal lung and breast cancer cell lines were previously published [6,15]. 11K data was obtained from prostate tumors, cell lines, and xenografts. 100K data was obtained from prostate tumors, gliomas, and lung cancer cell lines, along with paired normal DNA from (respectively) seminal vesicles, normal brain, and EBV-transformed lymphocytes. Tumor DNA was isolated from frozen tissue having >90% tumor content. DNA preparation and genotyping were performed according to manufacturer's instructions. Insufficient DNA was available in the case of one prostate tumor, four EBV-transformed lymphocytes, and the paired normal for one glioma. In these cases 20 ng of DNA was subjected to whole genome amplification [24] using the REPLI-g kit (Qiagen).

Reference normal samples. The heterozygosity rates for each SNP and the dependence information between the genotypes of neighboring SNPs were estimated from sets of normal samples; the haplotype correction was also performed against separate sets of normal samples (see Supplemental Methods). All these reference samples were from individuals unrelated to the tumor samples under evaluation. The estimated parameters are stored in genome information files available from the dChip website.

Observed LOH calls from paired normal and tumor samples. dChipSNP [12,25] was used to read CEL and TXT files containing the probe intensities and genotype calls (heterozygous AB, homozygous AA or BB, or missing genotype "No Call") [26]. The paired normal and tumor data were combined to make LOH calls for each SNP marker: loss (AB in normal, AA or BB in tumor), retention (AB in normal and tumor, or No Call in normal and AB in tumor), noninformative (AA or BB in normal, and the same genotype or No Call in tumor) or conflict (e.g. AA in normal, and AB or BB in tumor). A HMM was used to infer copy numbers at each SNP position from the probe level intensity data of the SNP arrays [15]. The positions of the SNP markers, genes, and cytobands were based on Affymetrix annotation files (www.affymetrix.com) and the UCSC human genome assembly (http://genome.ucsc.edu/).

RESULTS

A basic HMM for inferring LOH from unpaired tumor samples

The components of a HMM are the unobserved states, the observed measurements, the emission probabilities connecting these two, the transition probabilities between the unobserved states, and the initial probabilities of the states at the beginning of the chain (Figure 1). To infer LOH in unpaired tumor samples, we implemented a HMM with two unobserved states: loss (LOSS) and retention (RET) and the observed genotypes, reduced to homozygous (Hom; AA or BB), heterozygous (Het; AB), and "No Call". We conceptualize that the observed genotypes are generated by the unobserved LOH states according to the emission probabilities of the HMM.

Emission probabilities. For a SNP under the RET state, we observe Het calls with a probability equal to the heterozygosity rate of each SNP, which we estimated from normal samples (see Methods). For a SNP under the LOSS state, we always observe a Hom call unless a genotyping or SNP mapping error has occurred. Since genotyping errors occur at a rate < 0.01 [7], we set the emission probability of a Het call under the LOSS state to 0.01. The emission probability of the Hom call at a SNP is 1 minus the emission probability of the Het call at the SNP. A SNP with "No Call" could have had either an unobserved Hom or Het call, and is therefore emitted with a probability of 1 regardless of its underlying LOH state. As a result, a "No Call" does not bias the inference towards either LOSS or RET.

Initial probabilities. These probabilities (denoted by $P_0(\text{RET})$ and $P_0(\text{LOSS}) = 1 - P_0(\text{RET})$) specify the probabilities of RET and LOSS for the p-terminal marker on a chromosome. They also specify the probabilities of the RET and LOSS states for any marker, if no other information exists for that marker. Assuming Het markers are observed in regions of LOSS only as a result of genotyping or mapping errors, the observed proportion of Het markers in a tumor sample is ($P_0(\text{RET})$ *Average heterozygosity rate + $P_0(\text{LOSS})$ * SNP error rate). As the SNP error rate is small the 2nd term can be omitted. Therefore we estimate $P_0(\text{RET})$ by dividing the proportion of Het markers by the average heterozygosity rate of SNPs in the population.

Transition probabilities. These probabilities describe the dependence between the LOH states of adjacent markers. For any two adjacent SNP markers, we first define θ as the probability that the state of the first marker does not inform the state of the second (i.e. that the LOH state of the 2nd marker is distributed according to the Initial LOH state probabilities). Empirically, nearby

markers tend to have the same LOSS or RET state, while distant markers do not. To capture this observation we calculate θ using an increasing function $\theta = (1 - e^{-2d})$, where *d* is the physical distance (in the unit of 100 megabases (Mb) ~= 1 Morgan) between the two adjacent SNP markers. With probability 1 − θ, the two markers have the same LOH state. Therefore, the marker-specific transition probabilities of the 2nd marker being LOSS given the LOH state of the 1st marker are:

$P(\text{LOSS} \mid \text{LOSS}) = \theta \cdot P_0(\text{LOSS}) + (1 - \theta)$ and $P(\text{LOSS} \mid \text{RET}) = \theta \cdot P_0(\text{LOSS})$ (Equation 1)

The probability of RET at the 2nd marker is 1 minus these two probabilities. This transition probability model is the same as those used in the "Instability-Selection" model for LOH analysis [27,28], and is reminiscent of Haldane's map function in linkage analysis [22]. We used a fixed scaling of *d*, instead of estimating it as in the "Instability-Selection" model, but this does not affect the method performance (see below). In addition, the empirical transition frequencies estimated from observed LOH calls in paired normal and tumor samples agreed well with the transition probabilities estimated by this model (Figure 2).

Inferring LOH states. The HMM and these emission, initial, and transition probabilities specify the joint probability of the observed SNP genotypes and the unobserved LOH states in one chromosome of a sample. We applied the Forward-Backward algorithm [20] separately to each chromosome of each sample to obtain the LOSS probability for each SNP given all the genotype data on the chromosome. Alternatively, the Viterbi algorithm can be used; we found this gave similar LOH calls in 98.8% of SNPs (data not shown). LOSS and RET calls were made using the least stringent threshold: LOSS if the SNP has a probability of LOSS greater than 0.5 and RET otherwise.

An alternative inference method for HMM is the Baum-Welch algorithm [20], which estimates the model parameters together with unobserved LOH states by an iterative procedure. We chose not to use this algorithm as there are many parameters in the model (e.g. the transition probabilities depend both on the LOH states and on the distance between adjacent markers), but relatively few data points at each SNP position to estimate these parameters. This could lead the Baum-Welch algorithm to converge to local maxima when estimating optimal model parameters. Instead, we set biologically reasonable model parameters as above, with smooth transition probabilities that agree with the observed data (Figure 2). In addition, we show that the model inference accuracy is robust to the specified parameters in the initial, emission, and transition probabilities (see below).

The performance of the basic HMM. We compared tumor-only inferred LOH to the observed LOH calls determined by paired analysis of tumor and normal genotypes, using 10K SNP array data from autosomes of 14 lung and breast cancer and EBV-transformed normal cell line pairs [15] (Figure 3A). Here, 17,511 of 17,922 markers observed as LOSS in tumor/normal pairs were called LOSS in unpaired tumors by the HMM (for a sensitivity of 97.7%), and 15,962 of 16,364 markers that were observed as RET in tumor/normal pairs were called RET in unpaired tumors (for a specificity of 97.5%) (Supplemental Table 1A).

This initial analysis does not, however, account for the SNPs that are homozygous in both tumor and the paired normal, and thus are noninformative. A string of such homozygous SNPs may be falsely called LOSS in the HMM analyses of unpaired tumors, but not accounted for in the above

comparison of observed and inferred LOH states (the red arrows in Figure 3A point to two examples). To estimate the extent of such potentially falsely inferred LOH, we assigned an LOH state (LOSS or RET) to those noninformative markers for which the first informative marker on either side had the same LOH state. For example, a noninformative marker would be assigned a LOSS state if the nearest flanking informative markers were both in the LOSS state. In this analysis, the noninformative makers assigned a RET state were falsely inferred as LOSS at a rate of 6.8% (10K array) (Supplemental Table 1A). Not surprisingly, false inferences of RET were rare, occurring at a rate of 0.3%. Taking into account the noninformative markers in this way, the overall sensitivity remained high at 99.1%, but the specificity dropped to 94.3%. As an alternative approach to the use of flanking markers, we also inferred the LOH states of uninformative markers through the application of an HMM to the paired tumor/normal data, with nearly identical results (Supplemental Methods and Supplemental Table 2).

Linkage disequilibrium attenuates the performance of the basic HMM at high SNP density

With these methods in place, we next applied the basic HMM to 100K SNP array data from 2 prostate cancers and 2 lung cancer cell lines along with paired normal DNA, which were not included in the 10K dataset. Here, the number of noninformative regions inferred as LOSS increased significantly (Figure 3B). When noninformative marker status was assigned as above, many of these regions were deemed false regions of LOSS, and the specificity of the HMM decreased to 92.4% (Supplemental Table 1B). Furthermore, when 100K SNP array data derived from normal samples alone were analyzed, the basic HMM identified multiple regions of LOH that by definition are false (Figure 4A). We found that this occurred because, at high SNP densities regions of linkage disequilibrium (LD) are probed multiple times, resulting in strings of

homozygous SNPs. Specifically, if both parental chromosomes share the same haplotype, an extended stretch of homozygous genotypes will result. Therefore the assumption, inherent in the basic HMM, of independence between allele calls of adjacent or nearby SNPs becomes erroneous, leading to false inferences of LOH. An example is shown in Figure 4B, where the examination of an area of false LOH reveals the presence of a region of LD (dashed red box; also identified in the HapMap Project, www.hapmap.org).

HMM and haplotype correction that incorporate LD information

As indicated above, within a region of LD, the observed genotype of any marker depends not only on the underlying LOH state, but also on the genotypes of the adjacent markers (i.e. the two markers are dependent in genotype, indicated by the broken arrows in Figure 1). Here we account for many of these LD-induced SNP dependencies using an extension of the basic HMM (referred to herein as the Linkage Disequilibrium HMM or LD-HMM).

Expanded states and emission probabilities. We use the same observed Het and Hom genotypes of the tumor sample as in the basic HMM, but expand the unobserved LOH states for a SNP marker from the previous two states (LOSS or RET) to four states: Homozygous Loss (Hom LOSS), Heterozygous Loss (Het LOSS), Homozygous Retention (Hom RET) and Heterozygous Retention (Het RET). Here Hom and Het represent the SNP marker's genotype in the unobserved normal sample. For example, "Hom LOSS" represents that the SNP is homozygous in normal and LOH in tumor. The state "Hom LOSS", "Het LOSS" and "Hom RET" will result in homozygous genotype calls in the tumor unless genotyping or mapping error occurs, so the emission probability of the Hom genotype from these three states is set to (1 – SNP error rate).

The state "Het RET" will result in a heterozygous SNP call in the tumor unless a genotyping or mapping error happens, so the emission probability of the Hom genotype from this state is set to the SNP error rate. The emission probability of the Het genotype is 1 minus that of the Hom genotype.

Transition probabilities. The transition probabilities now reflect both the probability of a state change from RET to LOSS (LOH state), and a state change from Het to Hom (genotype state). We estimated genotype dependencies as the probability, for each SNP marker, of the next adjacent SNP marker towards q-arm being Hom (or Het), given the current SNP marker being Hom (or Het), in a reference set of normal samples (see Methods). We denoted these conditional probabilities for SNP *i* by $\mathrm{P}(U_{i+1} = \mathrm{Hom}|U_{\mathrm{i}} = \mathrm{Hom})$ and $\mathrm{P}(U_{i+1} = \mathrm{Het}|U_{\mathrm{i}} = \mathrm{Het})$. $\mathrm{P}(U_{i+1} = \mathrm{Het}|U_{\mathrm{i}} = \mathrm{Hom})$ and $\mathrm{P}(U_{i+1} = \mathrm{Hom}|U_{\mathrm{i}} = \mathrm{Het})$ are 1 minus the previous two probabilities respectively. When there were not enough data to estimate these probabilities at a marker, the SNP-specific heterozygosity rate was estimated from the reference set and used as the unconditional probabilities [e.g., replacing $\mathrm{P}(U_{i+1} = \mathrm{Het}|U_i = \mathrm{Hom})$ by $\mathrm{P}(U_{i+1} = \mathrm{Het})$, the heterozygosity rate of marker *i+1*]. Next, we built the transition probabilities by combining the above genotype dependence probabilities with the probability of an LOH state change. We denote the underlying LOH state of marker *i* by U_iV_i where U_i is either Hom or Het and V_i is either LOSS or RET. Suppose the current SNP *i* is in the "Hom LOSS" state while the next SNP *i+1* is in the "Het RET" state. For this to happen two independent events must occur: a homozygous genotype is followed by a heterozygous genotype in the normal with the probability $\mathrm{P}(U_{i+1} = \mathrm{Het}|U_i = \mathrm{Hom})$ estimated as above, and the LOH state changes from LOSS to RET in the tumor with the probability $P(V_{i+1} = \mathrm{RET} \mid V_i = \mathrm{LOSS})$ as specified in the transition probability

of the basic HMM. The transition probability from "Hom LOSS" to "Het RET" is then the product of these two probabilities. In general, the transition probability of going from LOH state U_iV_i to $U_{i+1}V_{i+1}$ is $P(U_{i+1}V_{i+1} \mid U_iV_i) = P(U_{i+1} \mid U_i)P(V_{i+1} \mid V_i)$.

Inferring LOH states. With the addition of the initial probabilities (which are the same as the basic HMM), the HMM parameters were fully specified and the Forward-Backward algorithm was used to obtain the probability of the LOH state being LOSS (either "Hom LOSS" or "Het LOSS") for every SNP, given all the observed SNP calls along one chromosome of a tumor sample. Application of the LD-HMM to the 100K dataset of normals, in place of the basic HMM, reduced the frequency of loss calls from 4.7% to 1.5% of markers (Figure 4C). Likewise, application of the LD-HMM to the 100K training dataset improved the specificity of LOSS calls from 92.2% to 97.4%, while only decreasing the sensitivity from 99.8% to 99.6% (Supplemental Table 1).

Empirical haplotype correction. We posited that the remaining regions of falsely inferred LOH resulted from three specific deficiencies of the LD-HMM. First, regions of LD might be present in a relatively small subset of patients [29]. Across the population as a whole, the genotypes of the neighboring SNPs within these LD regions correlate only weakly, and thus are not taken into account by the LD-HMM. Second, LD may happen between markers that are not immediately adjacent. Finally, in the LD-HMM, the dependency information among SNPs are estimated for the reduced genotype calls (Hom/Het) rather than from real genotypes. To try to address these concerns we also developed an empirical haplotype correction method, in which we applied a computational correction to the inferred LOH regions from either the basic or LD-HMMs (herein

referred to as HC-HMM and HC/LD-HMM, for the haplotype-corrected versions of the basic and LD-HMMs). For every putative LOH region called by HMM (LOH probability > 0.5 for all the SNP markers in the region but ≤ 0.5 for the SNPs at the boundaries of the region; containing mostly Hom SNP genotypes), we determine whether over 95% of the homozygous markers in this region in an unrelated normal reference sample are genotypically identical to the LOH region of the tumor sample. If this is the case, then the tumor sample is likely to share its haplotype structure with the reference sample in this region. Thus, homozygosity is likely due to LD rather than LOH, and the region is removed by setting the LOH probability of all the SNPs in the region to the LOH probability of the SNP marker just outside the bottom boundary of the region. This haplotype correction further improved specificity over the LD-HMM, in both the training 10K and 100K datasets, without significant loss of sensitivity (Figure 3, Supplemental Tables 1A and B).

The HC/LD-HMM infers LOH with high accuracy in a 100K validation dataset

To validate these results, we extended the analysis to a set of 100K data obtained from 2 lung cancer cell lines and 6 gliomas with paired normals, that had not been used in any of our prior analyses. Here, the sensitivity and specificity of the HC/LD-HMM were 98.7% and 99.3% respectively (Table 1 and Supplemental Table 1C). Compared to the basic HMM, the HC/LD-HMM led a greater than 8-fold reduction of potentially false LOH inferred at noninformative markers in the 100K data, but remained highly sensitive for real LOH events. Interestingly, once the haplotype block structure of the human genome is taken into account, the performance of HMM-based inferred LOH is better for 100K data than 10K data, presumably due to the denser SNP coverage of the 100K array.

LOH inference is robust to model parameter specifications

The methods described above rely on the empirical estimates of a number of the parameters used in the initial, emission, and transition probabilities of the HMM. To assess whether the tumor-only inference methods were unduly influenced by these estimates, we tested the performance of the basic and LD-HMMs as we varied these parameters. Specifically, the accuracy of the model results, as judged against observed LOH in the paired tumor/normal data, changed by less than 0.3% as the SNP error rate was varied from 0.1% to 1% (10K array). Moreover, when the SNP-specific heterozygosity rates were replaced by an average heterozygosity rate, that was varied from 0.1 to 0.5 (10K array) or from 0.1 to 0.27 (100K array), the accuracy of the model results changed by less than 5% and 0.5% respectively. We also found that varying the scaling factor *d* from 50 Mb to 200 Mb changed the LOH inferences of only 2% of SNP markers. These results suggest that the basic and LD-HMMs should be able to provide accurate LOH inferences in datasets that have different error rates, heterozygosity rates, or LOH-retention transition frequencies from the sample sets presented here.

Resolution of the HC/LD-HMM

The above analyses suggest that these methods are robust for inferring LOH on a per marker basis. We next asked whether the HC/LD-HMM was equally effective in detecting regions of LOH and whether detection of such regions was influenced by their size. To this end, we compared the ability of the tumor-only LOH analysis to identify LOH regions observed from comparing paired normal and tumor samples (Table 2). Here, we define a LOH region in the paired analysis as containing at least three LOH markers with any number of intervening

noninformative markers, and with boundaries defined in each direction by 2 consecutive retention markers. We considered such a region to have been "identified" by the tumor-only method, if that method inferred a probability of LOH > 0.5 for more than 90% of the SNP markers in the region. In the 100K datasets (both training and validation), the majority of regions of LOH observed in paired tumor/normal analysis are >3 Mb or are covered by at least 100 SNP markers, and >95% of these regions were identified using the unpaired analysis. Not surprisingly, smaller regions of LOH were detected less frequently. Overall, 80.8% of the regions of LOH identified in tumor/normal pairs were also identified in unmatched tumors in the 100K SNP data (Table 2). A similar analysis of the 10K data suggests higher sensitivity for smaller regions, apparently due to fewer such regions being identified by the tumor/normal paired analysis (Supplemental Table 3).

Integrating with copy number analysis to distinguish allelic imbalance

As mentioned in the introduction, LOH arises due to complete loss of one allele through hemizygous deletion (copy loss) or through gene duplication (copy neutral). On the other hand, heterozygous loci can erroneously be assigned a homozygous genotype in settings of allele specific amplification (allelic imbalance). This will occur whether or not LOH is determined using paired normals, and may present paradoxical results, with recurrently amplified oncogenes seen as potential TSGs. To address this issue we determined the copy number at each SNP locus using the probe level signal intensity data [15] and correlated the results with the LOH analysis. We found that among the observed LOH from normal/tumor pairs or the inferred LOH from unpaired tumors (using the basic HMM), about 70% of SNPs have copy number 2 (copy neutral LOH), 20% have copy number 1 (copy loss LOH), and 10% have copy number 3 or above

(amplification with possible allelic imbalance) (Figure 5). In contrast, among SNPs with observed retention from normal/tumor pairs or inferred retention from unpaired tumors, a lower percentage of markers have copy loss, and a higher percentage have amplifications (Figure 5). The combined LOH and copy number analysis can thus distinguish true LOH from those caused by amplification or allelic imbalance, which can be excluded from downstream LOH analysis. In addition, the copy number analysis can be used to distinguish LOH events caused by copy neutral gene conversion and copy number loss (Figure 5) [10,15]. In short, the vast majority of the regions of LOH detected using SNP arrays either by paired or unpaired analysis arises from copy neutral or copy loss events. Interestingly, the high frequency of copy-neutral LOH observed in these samples and others [3] suggests that LOH and copy number analyses provide independent sets of information pointing to TSGs.

Common LOH regions in a set of prostate cancer samples

Models of human cancer including xenografts and cell lines rarely are accompanied by paired normal samples. The utility of such models may be enhanced if we can ascertain the patterns of LOH in such models and relate them to those seen in actual human tumors. To this end, we next asked whether the HC/LD-HMM could detect regions of common LOH using 11K SNP array data from 34 prostate cell lines, xenografts, and metastases where the corresponding normal DNA was unavailable (Beroukhim et al, in preparation). We first scored each SNP by averaging the probability of LOH over all 34 samples (Figure 6, blue curves). The regions with the highest average probability of LOH correspond to known regions of frequent LOH, with several known and postulated TSGs lying in or near the regions with peak LOH scores (Figure 6, Supplemental

Table 4). These data suggest that the tumor-only LOH and copy number inference can be used to detect regions of true LOH where paired samples are not available.

DISCUSSION

We have developed an HMM-based method to infer the probability of LOH events from tumor samples without matched normals. The method utilizes several sources of information, including intermarker distances, SNP genotyping and mapping error rates, and haplotype information. LOH inferences using only tumor samples agree well with LOH patterns determined by analysis of tumor/normal pairs in two different array types (10K and 100K), three different tissue types (lung, glioma, and prostate), and in both cell lines and tumors, in test and in validation datasets. The inferences are robust to model parameter specifications. LOH is resolved to about 3 Mb or 100 SNPs in 100K array data. This method makes it feasible to use SNP array technology to map LOH in tumor samples for which normal DNA is unavailable. Given that genotyping paired normals samples constitutes up to half the cost of LOH mapping experiments, this method also makes it feasible to perform these experiments at a much lower cost per sample, at the expense of slightly reduced accuracy.

One advantage of a model-based approach over the existing tumor-only LOH inference methods [3,16] is its extensibility. The basic HMM was developed using average heterozygosity rates, but readily extended it to incorporate the SNP-specific heterozygosity rates and haplotype information as they became available. In addition, rather than making definitive calls the algorithm infers the probability of LOH at each marker of a sample. This SNP specific probability can then be used in further downstream analyses, such as identifying regions of shared LOH and sample clustering [5,25,27]. For example, a high probability of LOH across

many samples can indicate potential TSGs (Figure 6). The HMM approach can also be used to infer LOH probabilities for paired normal and tumor samples (see Supplemental Results), unifying the LOH analysis for paired tumor/normal and unpaired tumor samples.

At higher SNP densities, where the haplotype structure of the human genome becomes relevant, an approach that considers the dependence among multiple SNPs in a region of LD is necessary in addition to the LD-HMM. We used a haplotype correction that compared regions of inferred putative LOH to a set of reference normal samples to reduced false LOH inference. This method will work only if the reference samples have similar haplotypes to the tumor sample. As more data becomes available and the analyses are extended to other ethnic groups, it may become useful to utilize the data from the HapMap project to identify the haplotypes of each tumor prior to LOH analysis [30].

False designation of regions of LOH due to allelic imbalance may lead to paradoxical results, with recurrently amplified oncogenes seen as potential TSGs. SNP arrays, by providing signal intensity along with genotyping data, allow such regions to be identified. We can thus integrate these data to exclude regions of putative LOH with high copy numbers, as likely due to allelic imbalance. At the interpretive level, our finding that LOH is often copy-neutral suggests that LOH and copy loss should be considered independently when predicting the presence of a TSG, and may best be used in conjoined analyses.

The ability to identify regions of LOH in tumors without paired normal DNA allows LOH mapping in the many model systems lacking paired normal DNA, including cell lines and

xenografts. As such model systems are the platform for experiments aimed at understanding the biology of human tumors, it is critical that we understand their genetic relationship to real human tumors. As an example, among the prostate cancer samples, LOH at the *NKX3.1* locus is more prevalent among real tumors and xenografts than among cell lines, LOH at the *p53* locus is more prevalent among xenografts than among real tumors or cell lines, and LOH at the *Rb* locus is equally prevalent in all three groups (Figure 6). Larger sample numbers are required to see whether these differences are statistically significant. Such studies of the prevalence of regions of LOH across model systems compared to real tumors may indicate systematic faults in the ability of model systems to reflect *in vivo* cancer biology and guide the use and development of appropriate models based on genetic organization.

SNP array analysis of cancer genomes provides a single platform for copy number and LOH analysis. As these arrays move to higher resolution (500K), accounting for the haplotype structure of the human genome in the analysis of these data will be of greater import. The methods described herein, should be readily extensible to both the higher density arrays and to the increasingly detailed information describing the haplotype structure of the human genome. The software package, dChipSNP, is freely available at www.dchip.org.

ACKNOWLEDGMENTS

We thank L.J. Wei, M. Freedman and D. Altshuler for helpful discussions and J.G. Paez and C. Rosenow for training data. K. Pienta and the U. Michigan Prostate SPORE provided tumor tissues, and R. Vessella (U. Washington) and C. Sawyers (UCLA) provided prostate xenograft DNA. This work was supported by NIH grants P501062003, 1R01HG02341, R01CA109038 and

P20-CA96470 (KH, WHW and WRS), DOD grants PC040638 (RB) and W81XWH-04-1-0293, Friends of DFCI and Claudia Adams Barr Program (CL), Damon-Runyon Lilly Clinical Investigator Award (WRS), Tisch Family Foundation (MM and WRS), ASH Fellow Scholar Grant (EH), Flight Attendant Medical Research Institute (MM), and American Cancer Society (MM).

REFERENCES

1. Knudson AG (1971) Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68: 820-823.
2. Knudson AG (2001) Two genetic hits (more or less) to cancer. Nat Rev Cancer 1: 157-162.
3. Huang J, Wei W, Zhang J, Liu G, Bignell GR, et al. (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 1: 287-299.
4. McEvoy CR, Morley AA, Firgaira FA (2003) Evidence for whole chromosome 6 loss and duplication of the remaining chromosome in acute lymphoblastic leukemia. Genes Chromosomes Cancer 37: 321-325.
5. Girard L, Zochbauer-Muller S, Virmani AK, Gazdar AF, Minna JD (2000) Genome-wide allelotyping of lung cancer identifies new regions of allelic loss, differences between small cell lung cancer and non-small cell lung cancer, and loci clustering. Cancer Res 60: 4894-4906.
6. Janne PA, Li C, Zhao X, Girard L, Chen TH, et al. (2004) High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene 23: 2716-2726.
7. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, et al. (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21: 1233-1237.
8. Matsuzaki H, Dong S, Loi H, Di X, Liu G, et al. (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nature Methods 1: 109-111.
9. Matsuzaki H, Loi H, Dong S, Tsai YY, Fang J, et al. (2004) Parallel genotyping of over 10,000 SNPs using a one-primer assay on a high-density oligonucleotide array. Genome Res 14: 414-425.
10. Bignell GR, Huang J, Greshock J, Watt S, Butler A, et al. (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 14: 287-295.
11. Hoque MO, Lee J, Begum S, Yamashita K, Engles JM, et al. (2003) High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. Cancer Res 63: 5723-5726.
12. Lieberfarb ME, Lin M, Lechpammer M, Li C, Tanenbaum DM, et al. (2003) Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. Cancer Res 63: 4781-4785.
13. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, et al. (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat Biotechnol 18: 1001-1005.
14. Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, et al. (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res 10: 1126-1137.
15. Zhao X, Li C, Paez JG, Chin K, Janne PA, et al. (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64: 3060-3071.
16. Goldberg EK, Glendening JM, Karanjawala Z, Sridhar A, Walker GJ, et al. (2000) Localization of multiple melanoma tumor-suppressor genes on chromosome 11 by use of homozygosity mapping-of-deletions analysis. Am J Hum Genet 67: 417-431.

17. Wong KK, Tsang YT, Shen J, Cheng RS, Chang YM, et al. (2004) Allelic imbalance analysis by high-density single-nucleotide polymorphic allele (SNP) array with whole genome amplified DNA. Nucleic Acids Res 32: e69.
18. Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. J Mol Biol 268: 78-94.
19. Churchill GA (1989) Stochastic models for heterogeneous DNA sequences. Bull Math Biol 51: 79-94.
20. Durbin R, Eddy S, Krogh A, Mitchison G (1999) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press. 356 p.
21. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84: 2363-2367.
22. Lange K (2002) Mathematical and statistical methods for genetic analysis. New York: Springer-Verlag.
23. Fridlyand J, Snijders AM, Pinkel D, Albertson DG, Jain AN (2004) Hidden Markov models approach to the analysis of array CGH data. Journal of Multivariate Analysis 90: 132-153.
24. Paez JG, Lin M, Beroukhim R, Lee JC, Zhao X, et al. (2004) Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. Nucleic Acids Res 32: e71.
25. Lin M, Wei LJ, Sellers WR, Lieberfarb M, Wong WH, et al. (2004) dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data. Bioinformatics 20: 1233-1240.
26. Liu WM, Di X, Yang G, Matsuzaki H, Huang J, et al. (2003) Algorithms for large-scale genotyping microarrays. Bioinformatics 19: 2397-2403.
27. Miller BJ, Wang D, Krahe R, Wright FA (2003) Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides comparative evidence for multiple tumor suppressors and identifies novel candidate regions. Am J Hum Genet 73: 748-767.
28. Newton MA, Gould MN, Reznikoff CA, Haag JD (1998) On the statistical analysis of allelic-loss data. Stat Med 17: 1425-1445.
29. Liu N, Sawyer SL, Mukherjee N, Pakstis AJ, Kidd JR, et al. (2004) Haplotype block structures show significant variation among populations. Genet Epidemiol 27: 385-400.
30. Niu T (2004) Algorithms for inferring haplotypes. Genet Epidemiol 27: 334-347.

TABLES

Table 1. The number and proportion of SNP markers in the 100K validation dataset with LOSS or RET in tumor/normal pairs (LOH states were assigned to noninformative markers to agree with the nearest flanking informative markers), inferred as LOSS or RET by the basic HMM and HC/LD-HMM applied to the unpaired tumors.

		Basic HMM		HC/LD-HMM	
		LOSS	RET	LOSS	RET
Tumor/normal pairs	LOSS (171407)	170190 (99.3%)	1217 (0.7%)	169129 (98.7%)	2278 (1.3%)
	RET (702157)	42417 (6.0%)	659740 (94.0%)	4791 (0.7%)	697366 (99.3%)

Table 2. Percentage of LOH regions, identified in analysis of 100K data from tumor/normal pairs, that were also identified by the HC/LD-HMM applied to unpaired tumors, by size of region and the number of SNPs probed.

A.

Size of region (Mb)	Number of regions (per cent of total)	Number of informative SNPs (mean ± sd)*	Proportion identified by tumor only
≤ 1	54 (20.4%)	5.5 ± 4.4	40.7%
1 – 3	43 (16.2%)	10.2 ± 7.3	65.1%
3-10	46 (17.4%)	31 ± 23	91.3%
> 10	122 (46.0%)	437 ± 412	100%
All	265 (100%)	210 ± 350	80.8%

B.

Number of SNPs in region	Number of regions (per cent of total)	Number of informative SNPs (mean ± sd)*	Proportion identified by tumor only
1 – 40	48 (18.1%)	4.6 ± 2.6	25.0%
40 – 100	42 (15.8%)	9.3 ± 5.6	71.4%
100+	175 (66.0%)	314 ± 392	98.3%
All	265 (100%)	210 ± 350	80.8%

***"**sd" represents standard deviation

FIGURE LEGENDS

Figure 1. The elements comprising the HMM for LOH inference. Unobserved LOH states (LOSS or RET) of SNP markers generate observed genotype calls via emission probabilities. The solid arrows indicate the transition probabilities between LOH states, and the dashed arrows indicate LD-induced dependencies between consecutive SNP genotypes.

Figure 2. Comparison of empirically determined LOH transition probabilities (circles) to transition probabilities predicted by Equation 1 (black line) between retained loci (top panel) and loss loci (bottom panel).

Figure 3. Comparison of HMM-based LOH inferred from unpaired tumors to observed LOH based on tumor/normal pairs. A) Results from 10K SNP array data. Each column represents a sample, with SNP markers from chromosome 10 displayed from the p terminus (top) to the q terminus (bottom) (not all markers are displayed at this resolution). Tumor/normal observations (left panel) represent direct comparisons of tumor to normal genotypes. Here, SNP markers observed as having undergone LOH are indicated in blue, retention is shown in yellow, and noninformative SNPs are indicated in grey. Inferences from unpaired tumor data represent the probability of each SNP having undergone LOH, as made by the basic HMM (middle panel) and HC/LD-HMM (right panel). Here, a high probability of LOH (LOSS) is also indicated in blue, a high probability of retention (RET) is indicated in yellow, and indeterminate SNPs with an almost equal probability of either state are indicated in white. Occasionally, regions that are noninformative in the tumor/normal comparison are falsely inferred as LOH by the basic HMM in the unpaired data (red arrows); some of these false regions are corrected by the HC/LD-HMM

(green arrows). B) Results from 100K SNP array data. Panels are shown as in A. Data from chromosome 21 are shown to highlight the detection of false LOH in the analysis of unpaired tumor data, and are not representative of the frequency of true LOH events in this sample set. Almost all of regions falsely inferred as LOH by the basic HMM are correctly inferred by the HC/LD-HMM. The blue arrows indicate a region of true LOH, which is correctly identified by both the basic and HC/LD-HMM.

Figure 4. Accounting for linkage disequilibrium by the LD-HMM significantly reduces false LOH inferences from data obtained at high marker density. A) Inferences from the basic HMM applied to 100K SNP array data are shown for chromosome 4 in normal samples. Data are shown as in Figure 3. B) The genotypes of one region of falsely inferred LOH reveal a region of linkage disequilibrium (dashed red box), also identified by the HapMap project. The sample in column D contains one haplotype, the samples in columns E–K contain another haplotype, and the samples in columns A–C are heterozygous. C) Improved LOH inferences after application of the LD-HMM.

Figure 5. The proportion (y-axis) of the LOH (blue) or retention (red) markers observed from normal/tumor pairs in the 10K data, categorized by the inferred copy number at the same SNP markers (x-axis).

Figure 6. Inferred LOH in prostate cancer samples identifies regions of LOH known to be frequent in prostate cancer. The mean LOH probability across 34 prostate cancer samples is plotted along the left for all chromosomes. Peak regions of LOH are noted, and data from

chromosomes 8, 13, and 17 are highlighted on the right. These data are displayed as in Figure 3. Note that in this view, SNPs are visualized proportional to physical distance along the chromosome and most SNPs are not projected due to proximity to their neighbors. The red dotted lines indicate the approximate chromosomal positions of putative TSGs.

Figure 1

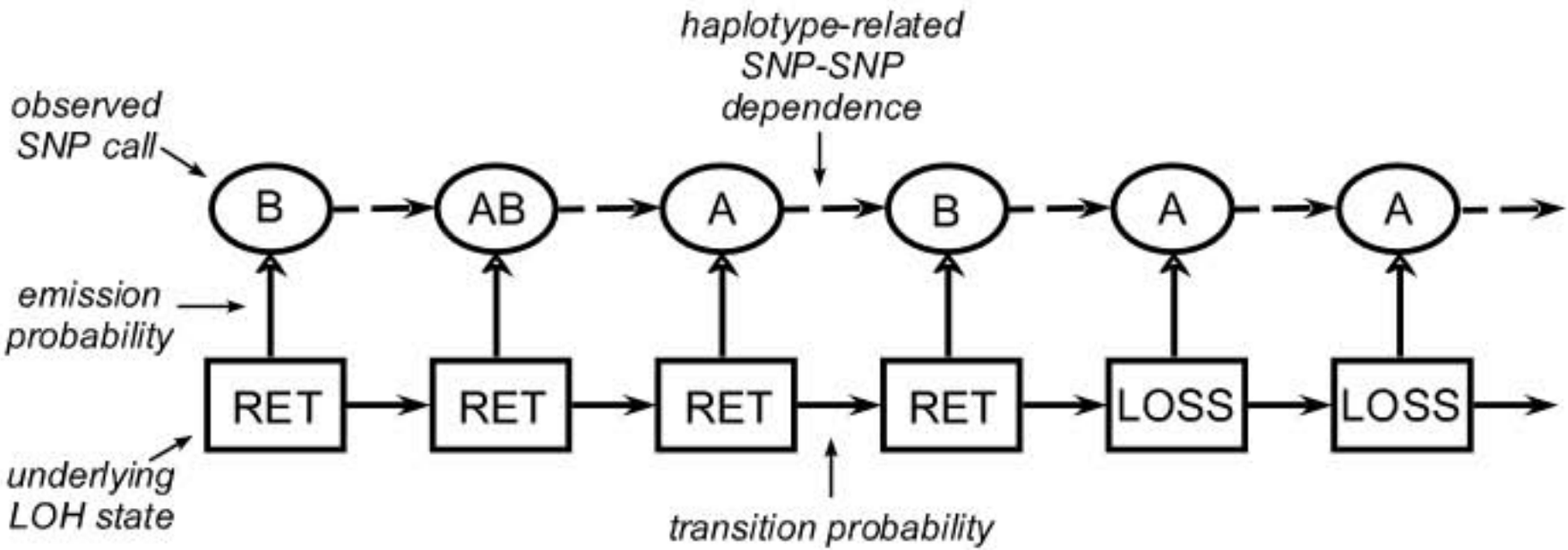

haplotype-related
SNP-SNP
dependence
observed
SNP call
B
AB
A
B
A
A
emission
probability
RET
RET
RET
RET
LOSS
LOSS
underlying
LOH state
transition probability

Figure 2

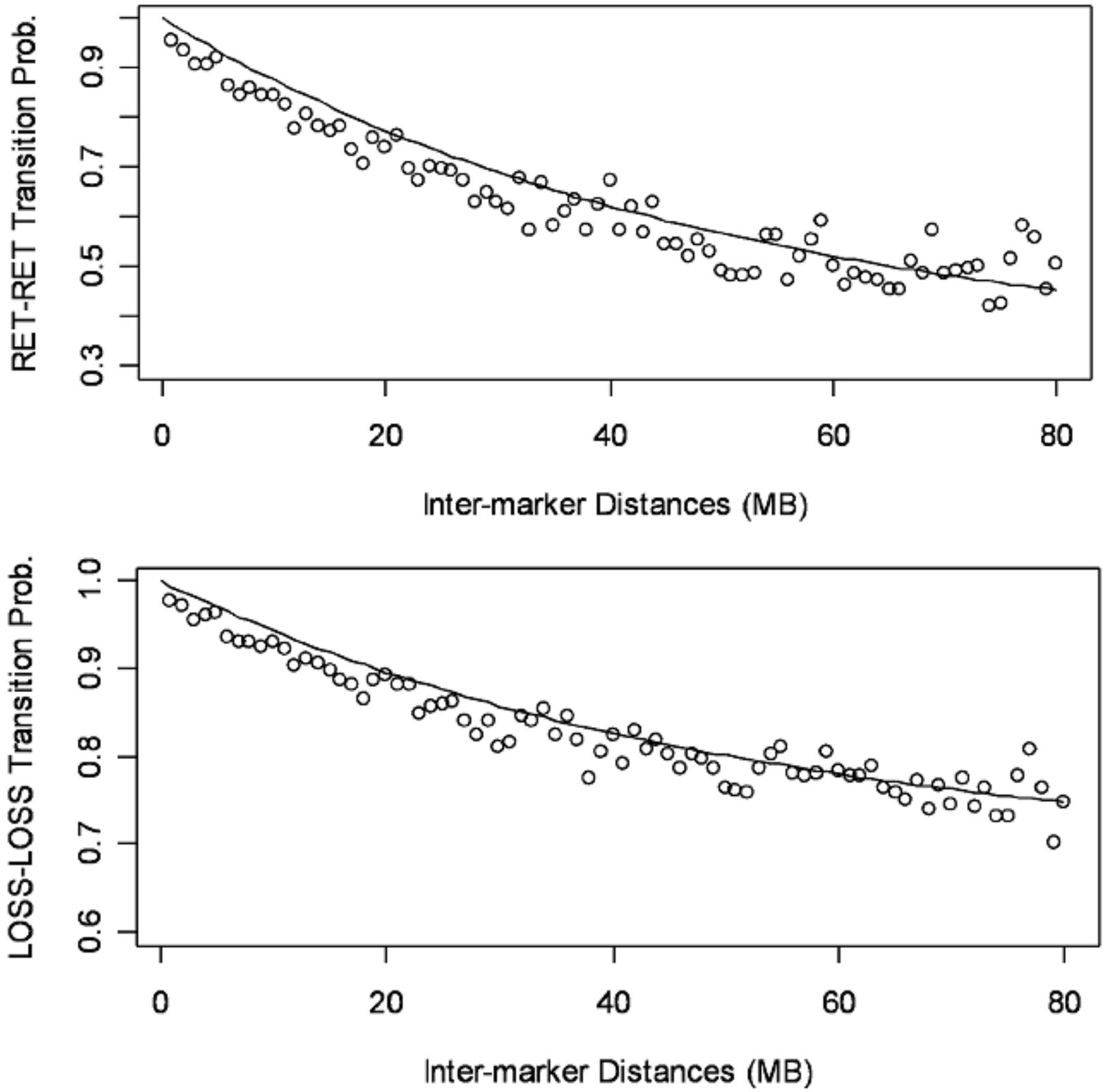

RET-RET Transition Prob.
0.3
0.5
0.7
0.9
0
20
40
60
80
Inter-marker Distances (MB)
LOSS-LOSS Transition Prob.
0.6
0.7
0.8
0.9
1.0
0
20
40
60
80
Inter-marker Distances (MB)

Figure 3

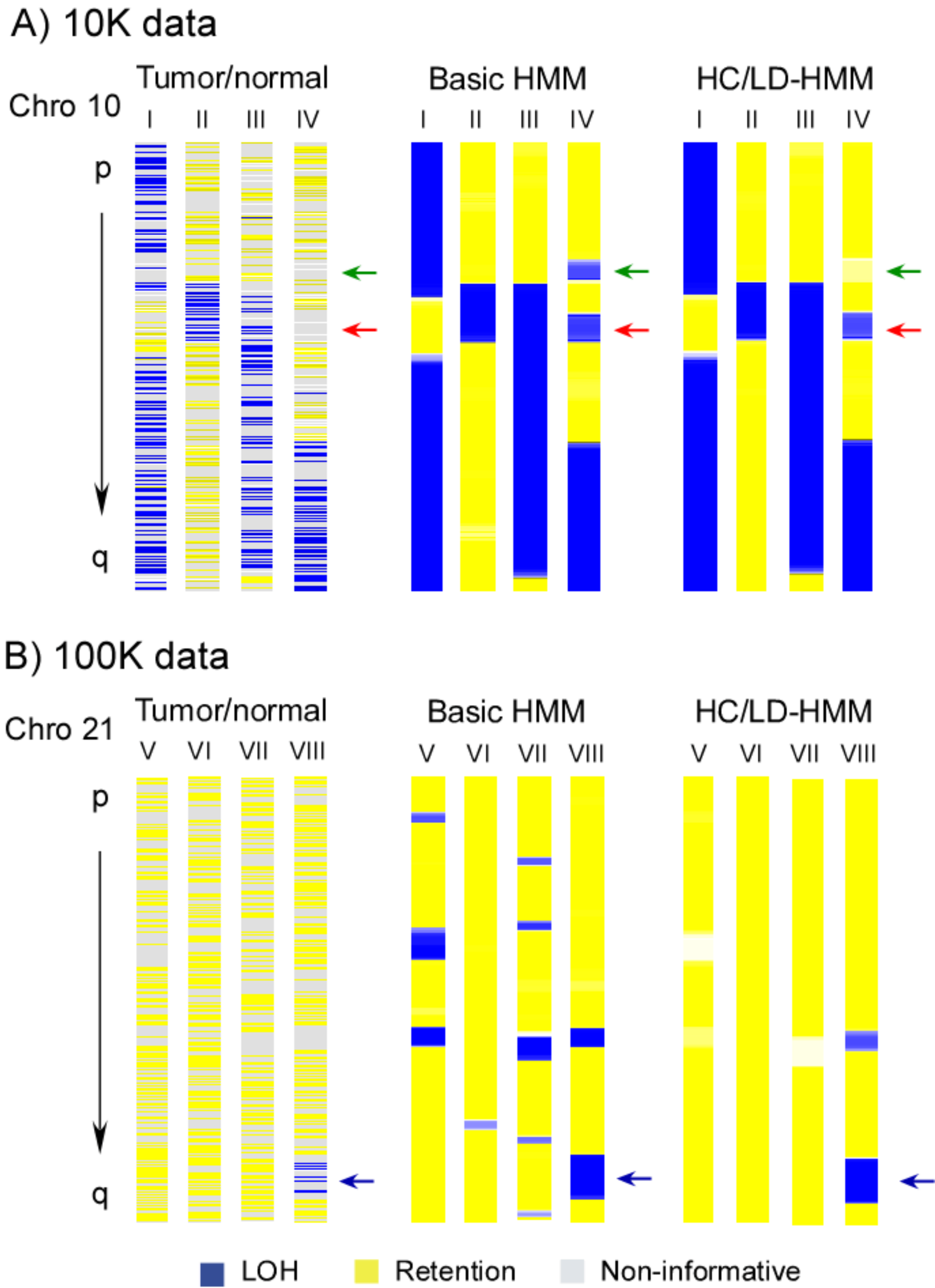

A) 10K data
Chro 10
Tumor/normal
Basic HMM
HC/LD-HMM
I
II
III
IV
p
q
B) 100K data
Chro 21
V
VI
VII
VIII
LOH
Retention
Non-informative

Figure 4

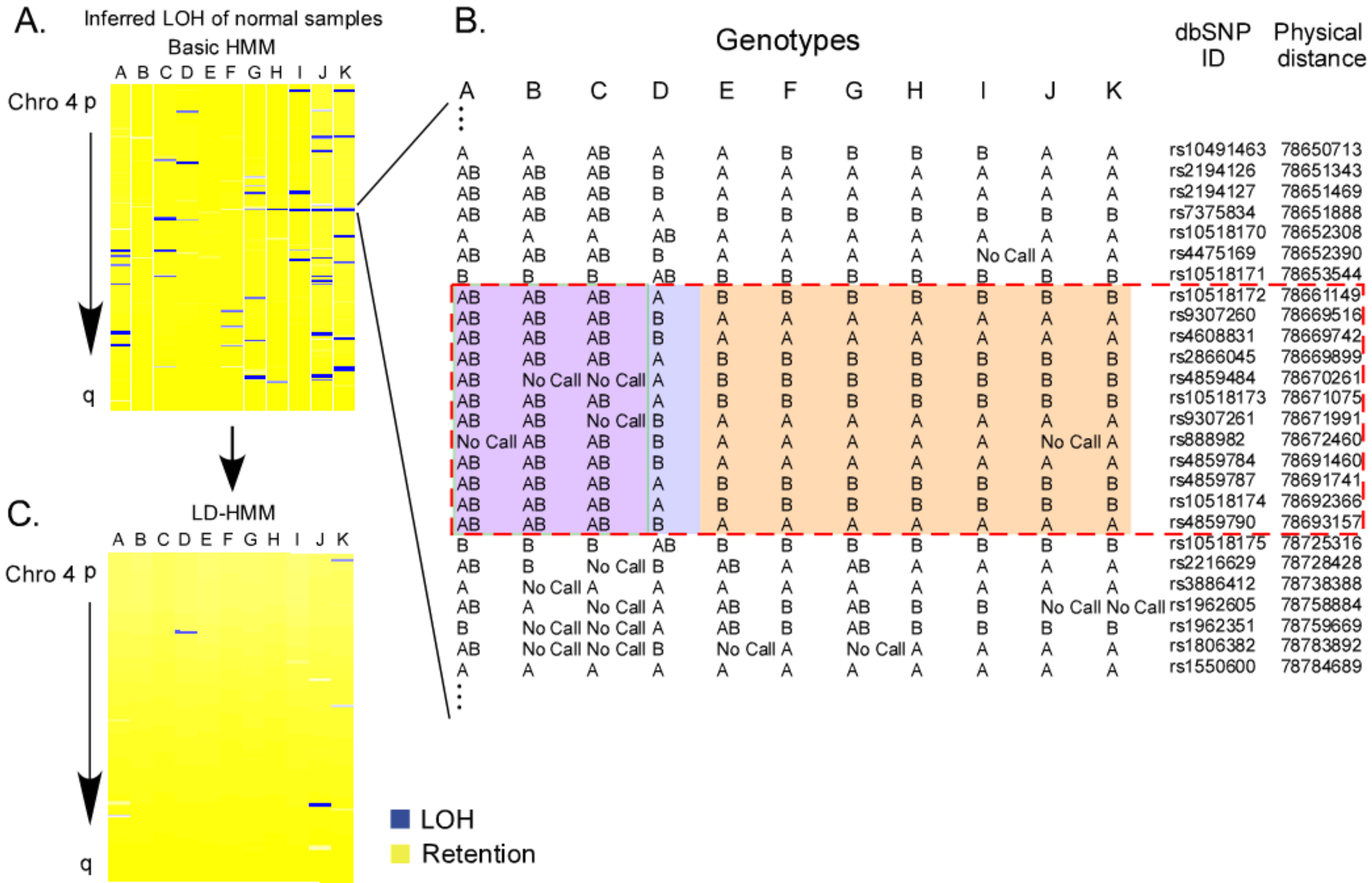

A.
Inferred LOH of normal samples
Basic HMM
A B C D E F G H I J K
Chro 4 p
q
B.
Genotypes
dbSNP ID
Physical distance
A B C D E F G H I J K
rs10491463 78650713
rs2194126 78651343
rs2194127 78651469
rs7375834 78651888
rs10518170 78652308
rs4475169 78652390
rs10518171 78653544
rs10518172 78661149
rs9307260 78669516
rs4608831 78669742
rs2866045 78669899
rs4859484 78670261
rs10518173 78671075
rs9307261 78671991
rs888982 78672460
rs4859784 78691460
rs4859787 78691741
rs10518174 78692366
rs4859790 78693157
rs10518175 78725316
rs2216629 78728428
rs3886412 78738388
rs1962605 78758884
rs1962351 78759669
rs1806382 78783892
rs1550600 78784689
C.
LD-HMM
A B C D E F G H I J K
Chro 4 p
q
LOH
Retention

Figure 5

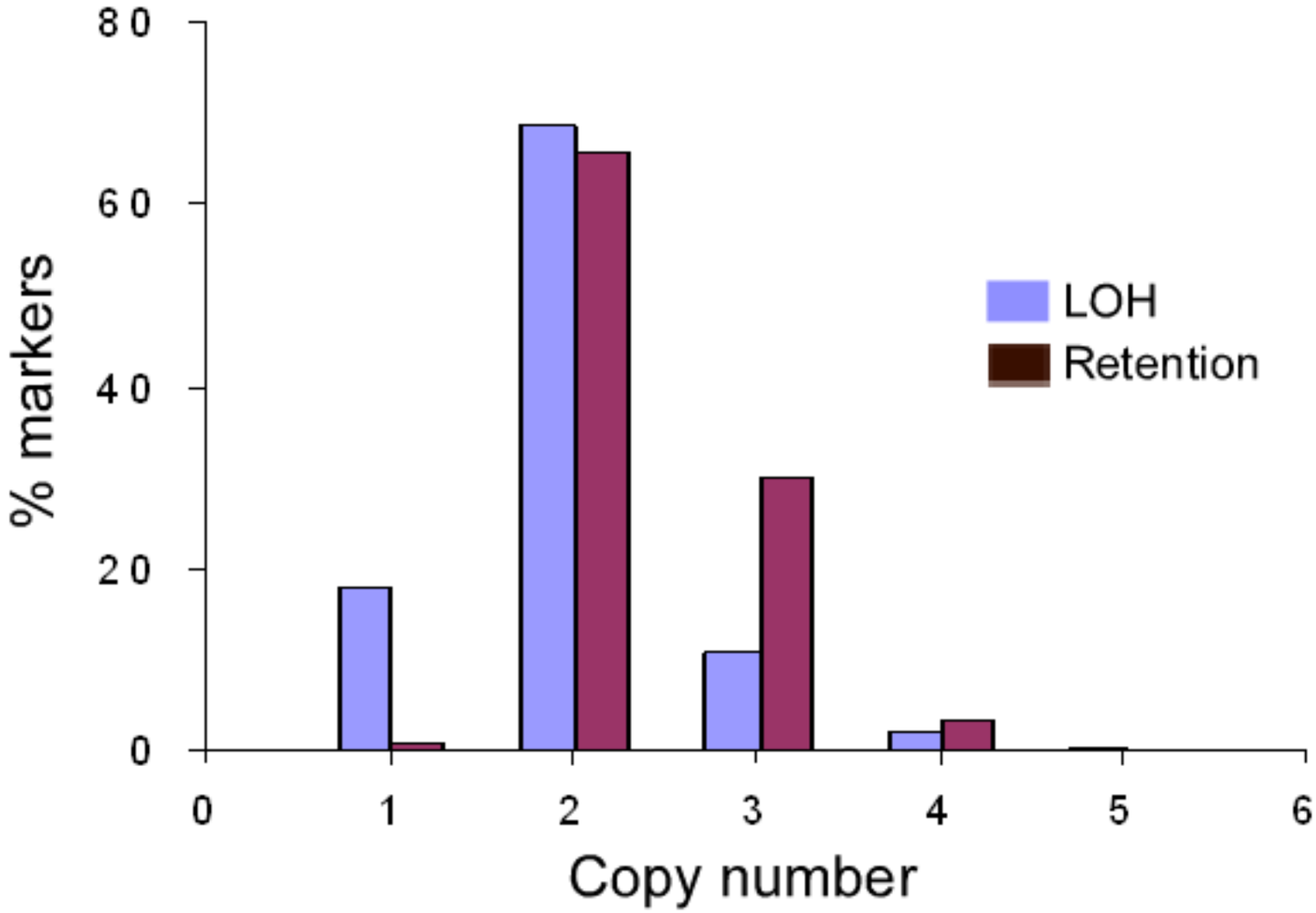
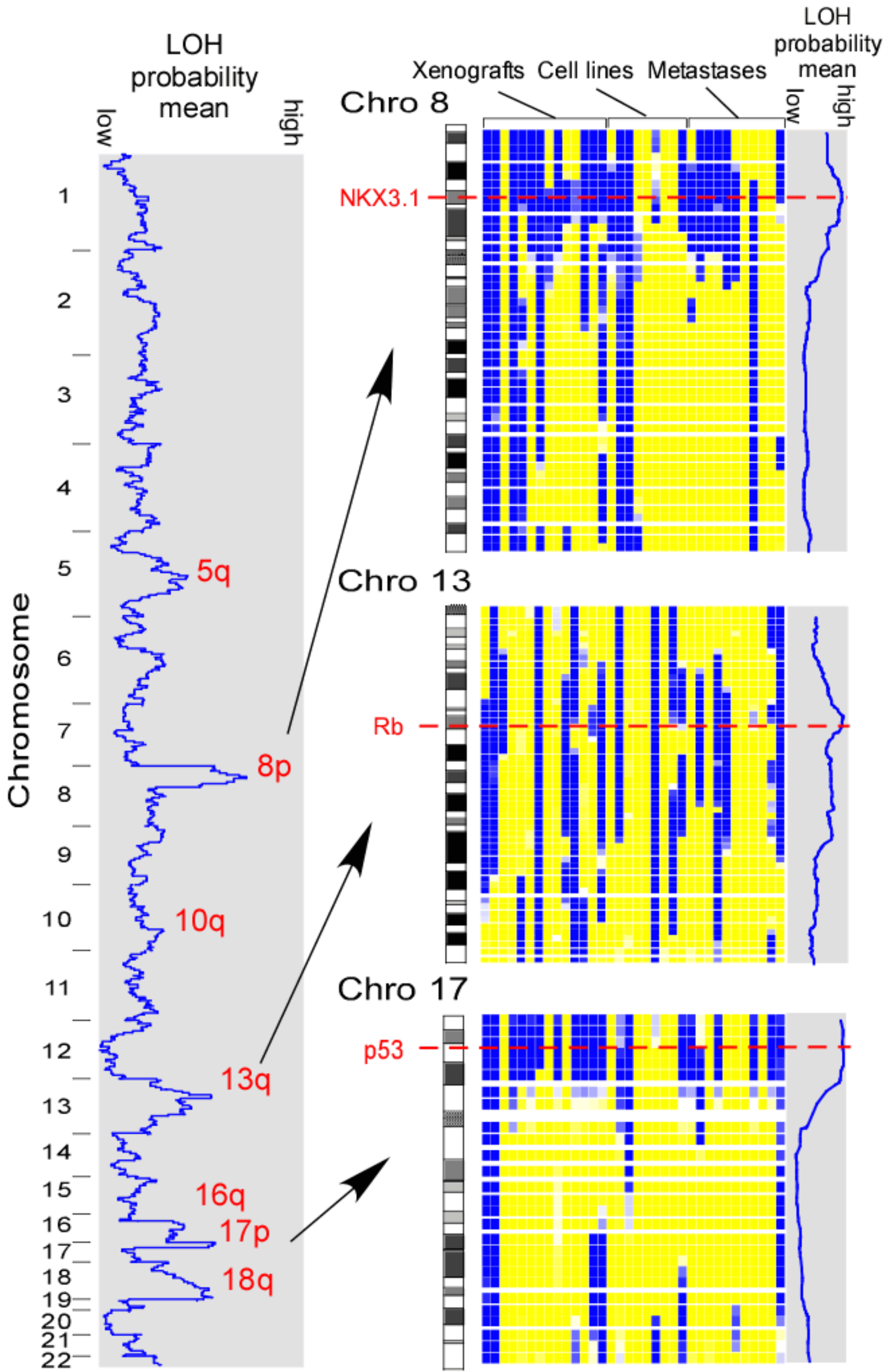


Figure 6


LOH probability mean
low
high
Chromosome
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
5q
8p
10q
13q
16q
17p
18q
Xenografts
Cell lines
Metastases
LOH probability mean
low
high
Chro 8
NKX3.1
Chro 13
Rb
Chro 17
p53